Kausale Inferenz in quasi-experimentellen Designs unter Verwendung von propensity score-Methoden

Dipl.-Psych. Marco Garling

28. April 2016

Dissertation zur Erlangung der Würde des Dr. phil.
an der Universität Hamburg
Institut für Psychologie
Psychologische Methodenlehre und Statistik

Erstgutachter: Prof. Dr. Martin Spieß
Zweitgutachter: Prof. Dr. Detlef Rhenius

Tag der Disputation: 27. April 2016

Promotionsprüfungsausschuss:

Vorsitz: Prof. Dr. Alexander Redlich

1. Dissertationsgutachter: Prof. Dr. Martin Spieß

2. Dissertationsgutachter: Prof. Dr. Detlef Rhenius

1. Disputationsgutachter: Prof. Dr. Matthias Burisch

2. Disputationsgutachter: Prof. Dr. Eva Bamberg

Inhaltsverzeichnis

Zusammenfassung

Die Messung des Effektes, der von einem applizierten treatment induziert wird und sich im Verhältnis zu einer Kontrollbedingung, in der dieses treatment nicht verabreicht wird, statistisch in einer beobachtbaren Differenz der Werte der abhängigen Variablen äußert, ist dem human- und sozialwissenschaftlichen Forschungsinteresse immanent. In einer Vielzahl empirischer Arbeiten, die aus den zugehörigen wissenschaftlichen Disziplinen hervorgehen, wird eine mögliche kausale Einordnung des gemessenen treatment-Effektes ersucht, die es in Folge dessen ermöglicht, die gemessene Varianz der abhängigen Variablen auf die Wirkung des treatments zu regressieren. Insbesondere das *randomisierte Experiment* als Forschungsdesign, das durch eine genaue Explikation der Experimentalbedingungen sowie durch eine zufällige Selektion der statistischen Einheiten in eine der Experimentalbedingungen charakterisiert ist, ermöglicht eine kausale Inferenz auf die Wirkung des treatments, da in Folge der Randomisierung jegliche Variablen, die in einem Zusammenhang mit der Messung des Effektes stehen, unabhängig von der Selektion sind. Damit ist in einem randomisierten Experiment die Annahme der *Unkonfundierung*, die eine Unabhängigkeit der response-Werte von der Selektion in die Experimentalbedingungen formuliert und zu Grunde gelegt werden muss, um den interessierenden *average treatment effect* anhand der Beobachtungen in statistischer Hinsicht identifizieren und konsistent schätzen zu können, erfüllt.

Es liegen dem randomisierten Experiment gegenüberstehend Forschungsdesigns, sog. *observational studies*, vor, in denen die statistischen Einheiten nicht randomisiert den Experimentalbedingungen zugeordnet werden können und denen einer möglichen Selbstselektion folgend eine Konfundierung, demnach eine Selektion in Abhängigkeit von den response-Werten, unterstellt werden muss. In Konsequenz lässt sich ohne weitere Annahmen der interessierende average treatment effect weder identifizieren noch anhand von Beobachtungen schätzen. Zu diesen Annahmen zählt das Modell eines *strongly ignorable treatment-assignment*, das eine von den beobachteten Kovariablen abhängige, probabilistische Selektion formuliert und damit die aufgeführte Konfundierung ersetzt durch eine von den Kovariablen abhängige Selektion in die Experimentalbedingungen; in Konsequenz können diese Kovariablen als Störvariablen wirkend sowohl die Selektionswahrscheinlichkeiten als auch die response-Werte beeinflussen. Dem strongly ignorable treatment-assignment folgend besitzt eine statistische Einheit eine feste, an den beobachteten Kovariablenwerten bedingte Selektionswahrscheinlichkeit, die als *propensity sco-*

re definiert ist und die es unter angenommener Gültigkeit der zugehörigen Modellannahmen ermöglicht, den average treatment effect zu identifizieren und konsistent zu schätzen, da bedingt am propensity score die Annahmen einer randomisierten Selektion greifen und das Design bedingt am propensity score unkonfundiert ist.

De facto ist der propensity score einer statistischen Einheit ein unbekannter Skalar und muss ausgehend von den beobachteten Kovariablenwerten geschätzt werden; üblicherweise wird hierzu in praxi ein logistisches Regressionsmodell spezifiziert, innerhalb dessen die Effekte der beobachteten Kovariablen auf die Wahrscheinlichkeit, zum treatment selegiert zu werden, geschätzt werden; die resultierenden geschätzten Selektionswahrscheinlichkeiten dienen damit als geschätzte propensity scores. Es lassen sich verschiedene Adjustierungsmethoden finden, den average treatment effect bedingt an den beobachteten propensity scores zu schätzen; hierzu zählen verschiedene Verfahren des propensity score-matchings, die propensity score-Stratifikation sowie verschiedene Methoden des propensity score-weightings.

In Anbetracht der geringen Anzahl an Veröffentlichungen, die einen systematischen Vergleich der verschiedenen propensity score-Methoden darlegen, werden in der ersten Simulationsstudie der vorliegenden Arbeit verschiedene vorgeschlagene propensity score-Methoden hinsichtlich ihrer Güte bei der Schätzung des interessierenden treatment-Effektes verglichen. Hierzu werden in mehreren Simulationsszenarien die Ergebnisse der Schätzungen des average treatment effects nach einer Adjustierung mit diversen matching-, Stratifikations- und weighting-Verfahren, die alle die geschätzten propensity scores nutzen, verglichen. Für einen erweiterten Vergleich werden zusätzlich Szenarien simuliert, die die Konsequenzen, die sich in Folge einer vom datengenerierenden Prozess abweichenden Spezifikation des logistischen Regressionsmodells bei der Schätzung des treatment-Effektes einstellen, aufweisen. Zu diesen Konsequenzen zählen unter Anderem, unabhängig von der gewählten propensity score-Methode, auffällige resultierende Verzerrungen bei der Schätzung des treatment-Effektes, die sich einem fehlspezifizierten logistischen Regressionsmodell zur Schätzung des unbekannten propensity scores folgend einstellen.

Ausgehend von diesen Ergebnissen wird im zweiten Teil dieser Arbeit ein non-parametrisches Verfahren zur Schätzung des unbekannten propensity scores vorgeschlagen, das sich auf die Klasse der stetigen Kovariablen mit einer zugehörigen log-konkaven Dichte anwenden lässt. Hierzu wird zunächst theoretisch durch die Anwendung des Satz von Bayes der von Cule, Samworth und Stewart (2010) vorgeschlagene non-parametrische Dichteschätzer mit dem un-

bekannten propensity score in Verbindung gebracht; dem folgend wird die vorgeschlagene propensity score-Schätzung in mehreren Simulationsszenarien bei der adjustierten Schätzung des treatment-Effektes evaluiert.

Im dritten Teil dieser Arbeit wird eine Adjustierungsmöglichkeit zur Schätzung des average treatment effects vorgeschlagen, die sich in Szenarien mit latenten konfundierenden Störvariablen anwenden lässt und auf den geschätzten propensity scores beruht: Ausgangspunkt dieses Vorschlags ist das Vorliegen einer latenten Kovariablen, die sich durch manifest erhobene Indikatorvariablen messfehlerbehaftet operationalisieren lässt und in ihrer Wirkung als Störvariable einerseits die response-Werte, andererseits die Selektionswahrscheinlichkeit einer statistischen Einheit beeinflusst. Es wird zunächst in theoretischer Hinsicht der Vorschlag zur Schätzung des propensity scores, geschätzt aus den resultierenden Hauptkomponenten der Indikatorvariablen, vorgestellt und in mehreren Simulationsszenarien hinsichtlich der Güte der Schätzung des treatment-Effektes evaluiert. Diese Evaluierung umfasst zusätzlich einen Vergleich mit den Ergebnissen der Schätzung des treatment-Effektes nach Adjustierung innerhalb eines *Strukturgleichungsmodells*, das in praxi üblicherweise für dergestalte Adjustierungen genutzt wird.

1 Theoretischer Hintergrund

1.1 Einführung: Der Begriff des kausalen Effektes

Eine Klärung des Begriffes *Kausalität* ist für das human- und sozialwissenschaftliche Forschungsinteresse unerlässlich, da jedwede durchgeführte Intervention - beispielsweise eine psychotherapeutische Maßnahme oder eine medizinische Therapie - verbunden wird mit der Annahme, dass diese ursächlich eine Veränderung des vorangegangenen, zumeist unerwünschten Zustandes evoziert. Petersen, Sinisi und van der Laan (2006, S.276) verdichten:

> „Research may be aimed at understanding the mechanism by which a treatment or exposure affects disease.“

Dabei beschränkt sich der praktische Nutzen einer gesicherten kausalen Beziehung zweier Variablen nicht nur auf psychologische oder medizinische Maßnahmen, sondern kann in andere Forschungsfelder und Fragestellungen implementiert werden, beispielsweise bei der Evaluierung makroökonomischer Maßnahmen (z.B., Blundell & Costa-Dias, 2002; Dehejia & Wahba, 2002; Heckman, Ichimura & Todd, 1997), oder bei pädagogischen Fragestellungen (Morgan, 2001). Sämtlich aufgeführten Beispielen gemeinsam ist die statistische Evaluierung der Frage, inwieweit eine ergriffene Maßnahme (eine ABM, eine Psychotherapie oder ein Schulwechsel auf eine katholische Schule) einen messbaren Effekt evoziert; zumeist handelt es sich bei diesem Effekt um eine Verbesserung des vorherigen Zustandes oder um einen messbaren Unterschied zu einem alternativen Zustand, der sich ohne diese Maßnahme auszeichnet.

Allgemein gültig ist die Feststellung einer kausalen Beziehung zwischen zwei Variablen für die praxisbezogene Forschung von Interesse, da - unter der Bedingung, dass die Wirkungsweise einer Variable S_i auf eine Variable Y_i eindeutig festgestellt worden ist - eine Prävention oder Veränderung möglich ist, indem eben nur unter dieser Bedingung die Ursache (eine Ausprägung $S_i = s_i$) für den unerwünschten Zustand ($Y_i = y_i$) verändert werden kann: Nur wenn bekannt ist, dass ein bestimmter Umweltfaktor ($S_i = s_i$) einen unerwünschten Zustand oder eine Krankheit hervorruft ($Y_i = y_i$), kann dem, beispielsweise therapeutisch oder präventiv, entgegen gewirkt werden durch eine Veränderung des krankheitsinduzierenden Umweltzustandes; statistisch demnach durch Herbeiführen eines alternativen $S_i = s_i'$ (Hill, 1965).

Grundlegend lassen sich empirisch gewonnene Erkenntnisse über den Zusammenhang zweier Variablen in *descriptions* und *explanations* kategorisieren: Für funktionale Aussagen, wie sie

Interventionen oder Behandlungen zu Grunde liegen, sind lediglich explanations brauchbar. Nur auf ihrer Basis können Funktionalitäten zwischen Variablen beschrieben werden und zur Modifikation von, zumeist malignen, Bedingungen genutzt werden. Wold (1956, S. 29) verdichtet entsprechend:

> „A frequent situation is that description serves to maintain some *modus vivendi* (...), whereas explanation serves the purpose of *reform* (...). In other words, description is employed as an aid in the human *adjustment* to given conditions, while explanation is a vehicle for *ascendancy* over the environment.“

Folgende **Definition** wird für die vorliegende Arbeit Gültigkeit haben (z.B., Holland, 1986; Holland & Rubin, 1988): Statistisch handelt es sich bei dem Begriff der Kausalitität um eine Gleichsetzung mit einem empirisch erfassbaren Effekt, der aus einer asymmetrischen Beziehung zweier Variablen hervorgeht und sich niederschlägt in einer Veränderung der einen Variablen (*beoabachtete abhängige Variable*, Y_i), insofern zeitlich eine Variation der anderen Variablen (*unabhängige Variable*, S_i) vorangeht.

Explizit kann ein kausaler Effekt nur in Verbindung mit einer gezielt durchgeführten Intervention (folgend: *treatment*) - stellvertretend für eine (Be-)Handlung, deren Wirkung empirisch erfasst werden soll - gemessen werden; es handelt sich bei einer *kausalen Inferenz* stets um die Messung eines Effekts, der durch diese Intervention evoziert wird und der in einer messbaren Größe - Y_i- eine Veränderung hervorruft, insofern im Verhältnis zu einem Ausgangspunkt jenes treatment durchgeführt wird (*vorher-nachher-Messung*) bzw. insofern zwei verschiedene Handlungen (im einfachsten Falle: *treatment* im Verhältnis zu dem Zustand *kein treatment*) verglichen werden. Die auf der abhängigen Variablen messbare Veränderung darf, um als kausal eingeordnet zu werden, lediglich der Variation der unabhängigen Variablen folgend systematisch auftreten[1] und muss bei Wirkung des treatments eine Differenz im Verhältnis zu der Kontrollbedingung, welche die (Be-)Handlung nicht erfahren hat, evozieren (Cox, 1992). Die entscheidende Annahme, die an die philosophischen Überlegungen, z.B. von Hume (1739), anknüpft: Die Ursache - in diesem Falle die Durchführung des treatments - hat zeitlich vor der Beobachtung der Wirkung - die Messung der Veränderung in der abhängigen Variablen - zu liegen.

[1] Damit wird der noch zu klärenden Anforderung Genüge geleistet, dass der gemessene Effekt nicht auf weitere Varianzquellen regressiert werden kann.

Dem folgend vorzustellenden *Rubin-Modell der kausalen Inferenz* nach lässt sich ein kausaler Effekt, der einem wirkungsvollen treatment folgt, empirisch erfassen, wenn an einer statistischen Einheit, folgend als i indiziert, die ihr zugehörigen Werte der abhängigen Variablen Y_i sowohl unter der Bedingung, in der das treatment nicht appliziert wird, als auch unter der Bedingung des durchgeführten treatments gemessen werden könnten und eine Differenz zwischen den resultierenden Werten der abhängigen Variablen festzustellen wäre (z.B., Holland, 1986; Rubin, 1974). Der Exposition in die Kontrollbedingung folgend nimmt die Variable Y_i den Wert $y_{i,c}$, der Exposition in das treatment folgend den Wert $y_{i,t}$ an, der den Messwert nach Durchführung des treatments indiziert. Die messbare Differenz zwischen den beiden Werten, so wird es durch t und c als Index impliziert, wird ausschließlich durch die Manipulation der unabhängigen Variablen, mit den Ausprägungen „treatment“ und „Kontrollbedingung“, hervorgerufen; die Beziehung zwischen S_i und Y_i gilt als kausal, wenn jene Differenz der Messwerte nur in Folge der Applikation des treatments beobachtet wird.

1.2 Explikation: Kausaler Effekt

Grundlegend können kausale Effekte nur in Studiendesigns erfasst werden, die als *kausale Agenten* (Holland & Rubin, 1988) im einfachsten Fall eine Kontrollbedingung c und ein treatment t definieren; folgend zusammengefasst zu der Menge $K := \{c, t\}$, die dergestalt ist, als dass es sich bei dem Element $t \in K$ um eine Handlung handelt, die einer Einheit i zukommen kann, während $c \in K$ als Kontrollbedingung so definiert ist, das es als Studienbedingung eine Auslassung dieser Handlung beinhaltet. Damit erfüllt die Definition der Menge K die fundamentale Voraussetzung für einen kausalen Rückschluss, die als *potential exposability*-Annahme in der Literatur zu finden ist (z.B., Holland, 1986; Imbens & Rubin, 2012): i muss stets allen $k \in K$ potentiell ausgesetzt werden können; insofern t als eine (Be-)Handlung definiert ist, gilt diese Annahme als erfüllt; Charakteristika hingegen verstoßen gegen diese Annahme.

Nach obiger Explikation ist der *individuelle kausale Effekt*, folgend τ_i, der einem wirkungsvollen t folgt, definiert als die Differenz der response-Werte $y_{i,t}$ und $y_{i,c}$, die sich messen ließe, wenn die statistische Einheit i, zufällig einer interessierenden, finiten Grundgesamtheit U ent-

nommen[2], sowohl t als auch c expositioniert wird, formal:

$$\tau_i := y_{i,t} - y_{i,c}. \tag{1}$$

Mit dieser Definition einhergehend besitzt i sowohl einen festen response unterhalb der Kontrollbedingung, $y_{i,c}$, wie auch unterhalb des treatment, $y_{i,t}$. Beide response-Werte sind an i *potentiell* mit Exposition in beide Bedingungen beobachtbar und geben, evoziert durch die Durchführung der Experimentalbedingungen, unmittelbar Aufschluss über die individuelle Wirkung von t[3].

An dieser Stelle knüpft das dargelegte Modell an das philosophische *ceteris paribus*-Prinzip an, da eine Beobachtung der Werte $y_{i,t}$ und $y_{i,c}$ inhaltlich mit der Feststellung, wie i einerseits auf t, andererseits auf die Auslassung des treatments in c reagiert, einhergehen würde (Morgan & Winship, 2007). Es lässt sich unmittelbar ein Bezug zur potential exposability-Annahme herstellen: das ceteris paribus-Prinzip kann nur sinnvoll angewandt werden, wenn die Experimentalbedingungen so definiert sind, dass i hätte allen potentiell ausgesetzt werden können. De facto kann i, wie folgend als *Fundamentalproblem der kausalen Inferenz* expliziert wird, nicht allen $k \in K$ expositioniert werden, so dass ein kausaler Rückschluss stets ein *missing-data*-Problem darstellt und folglich verbunden ist mit der Frage: „wie hätte i auf die ausgelassene Bedingung reagiert?“.

Es ist dauerhaft vorauszusetzen, dass die Exposition von i zu den Experimentalbedingungen zeitlich vor der Messung des zu dem $k \in K$ gehörigen Wertes $y_{i,k}$ gelegen ist, da nur dementsprechend an das Ursache-Wirkungs-Prinzip der kausalen Inferenz angeknüpft wird, welches davon ausgeht, dass die Ursache (in diesem Falle die Durchführung von t und c) und die Wirkung (Messung von (1)) in einer eindeutigen zeitlichen Sequenz zueinander stehen. Dorn (1953, S. 677) hebt diese Annahme hervor:

> „The most general basis for belief in the cause and effect relationship of events is the observation that they are sequentially related in time. The first event is then thought to be the cause of the second“.

[2]Als Grundgesamtheit U wird im Folgenden diejenige Menge der statistischen Einheiten verstanden, über die ausgehend von dem Kontext der Evaluationsstudie eine Aussage getroffen werden soll. Demnach ist U als diejenige statistische Masse zu verstehen, die sich in sachlicher, zeitlicher und räumlicher Hinsicht im Rahmen einer Studie abgrenzen lässt, welche z.B. von t profitieren würde bzw. für die generalisiert die Wirkung von t nachgewiesen werden soll.

[3]Sog. *stable unit treatment value assumption* oder SUTVA (z.B., Holland, 1986; Imai, King & Stuart, 2008; Imbens & Rubin, 2012)

In diesem Zusammenhang können die potentiellen response-Werte für i unter den beiden kausalen agents, $y_{i,t}$ und $y_{i,c}$, nur ein potentielles Maß für die Wirkung von t darstellen, wenn diese sog. „post exposure“-Werte sind (Holland, 1986), demnach der Durchführung der Experimentalbedingungen folgend beobachtet werden.

1.3 “Fundamentalproblem der kausalen Inferenz“ und Selektionen

Zusammenfassend gilt, dass i durch das Wertepaar $(y_{i,t}, y_{i,c})$ charakterisiert wird, welches potentiell bei einer simultanen Exposition zu t und c beobachtbar wäre - eine Messung des in (1) definierten kausalen Effektes, welcher den fundamental interessierenden Effekt darstellt, wäre eine „unit-level causal inference“ (Holland & Rubin, 1988).

Einerseits stellt eine individuelle kausale Inferenz die stärkste Form eines möglichen kausalen Rückschlusses dar, da sie unmittelbar Evidenz über die Wirkung von t gibt, andererseits lässt sie sich praktisch nicht realisieren, da i in praxi ausschließlich unterhalb von t oder c beobachtet wird: Explizit in den Humanwissenschaften muss davon ausgegangen werden, dass i nach entsprechender Exposition in eines der $k \in K$ nicht mehr unabhängig auf die anderen Versuchsbedingungen reagieren wird - es müssen sog. *carry over*-Effekte unterstellt werden (z.B., Holland, 1986; Rubin, 1974). In praxi folgt damit, dass i einer der Versuchsbedingungen zugewiesen und ausschließlich unterhalb dieser Bedingung beobachtet wird; somit lässt sich faktisch der in (1) definierte Effekt nicht beobachten, da einer der beiden Werte des Paares $(y_{i,t}, y_{i,c})$ nicht gemessen werden kann, sondern nur *kontrafaktisch*, das heißt als existent angenommen, aber unbeobachtbar, vorliegt (sog. *Fundamentalproblem der kausalen Inferenz*). Eine Inferenz auf die Wirkung des treatments erscheint damit zunächst nicht möglich.

Wie in folgenden Abschnitten expliziert wird, variiert die notwendige Selektion der Einheit i in eine der Experimentalbedingungen nach Studiendesign, jedoch wird dauerhaft vorausgesetzt, dass diese Selektion das Resultat eines, wenn auch nicht notwendigerweise bekannten, Zufallsversuchs ist. Damit lässt sich als Selektionsindikator die Zufallsvariable S_i definieren, deren Realisierung $S_i = s_i$ die Selektion von i zu $c \in K$ bzw. $t \in K$ numerisch mit $s_i \in \{0, 1\}$ indiziert, wobei fortlaufend:

$$S_i = \begin{cases} 1, & \text{wenn } i \text{ in } t \\ 0, & \text{wenn } i \text{ in } c. \end{cases}$$

Offenkundig entscheidet die Zuweisung von i zu t oder c darüber, welcher der beiden potentiellen

Werte beobachtet wird, da i folglich der ausgelassenen Bedingung nicht mehr ausgesetzt werden kann. Das Fundamentalproblem der kausalen Inferenz formuliert damit die faktisch unmögliche Beobachtung des Wertepaares $(y_{i,c}, y_{i,t})$ in Abhängigkeit von der notwendigen Selektion; in Folge lässt sich folgend für den beobachteten response y_i festhalten:

$$y_i = y_{i,t} \cdot s_i + y_{i,c} \cdot (1 - s_i) = y_{i,c} + s_i \cdot (y_{i,t} - y_{i,c}) = y_{i,c} + s_i \cdot \tau_i \tag{2}$$

Wie an späterer Stelle gezeigt wird, führt die Restriktion, dass die Selektion von i zufällig ist, zu Lösungsmöglichkeiten dieses Fundamentalproblems.

1.4 Designs zur Messung kausaler Effekte

Randomisierte Experimente

Besonders in *randomisierten Experimenten* ist die Erfassung eines kausalen Effektes *intern hoch valide* (Campbell, 1957), da zum Einen der Definition der Menge K folgend eine kontrollierte Varianz der unabhängigen Variablen vorliegt, zum Anderen die Selektion von i zu t oder c durch einen bekannten Zufallsmechanismus, der die Realisierung $S_i = s_i$ generiert, beispielsweise durch einen Münzwurf, vollzogen wird (Rosenbaum, 1995).

Randomisierte Experimente zeichnen sich grundlegend dadurch aus, dass *a-priori*, d.h. vor Durchführung, festgelegt wird, wie wahrscheinlich es für i ist, dass diese das treatment erhält; damit basiert die Selektion von i zu t auf einer positiven, festen und bekannten Wahrscheinlichkeit, $\Pr(S_i = 1)$, so dass: $0 < \Pr(S_i = 1) < 1, \forall i \in U$. Dem folgend ist die resultierende Varianz in S_i kontrolliert und kalkulierbar, da der zu Grunde liegende Zufallsprozess, der die Selektion bedingt, von dem Versuchsleiter durchgeführt wird. Wie an späterer Stelle gezeigt wird, schließt dies eine Kenntnis über das komplette *treatment-assignment* ein: Mit Festlegung der Selektionswahrscheinlichkeit $\Pr(S_i = 1)$ sind sämtlich mögliche Selektionsindikationen der N Einheiten in beide Experimentalbedingungen, folgend als Vektor $\vec{S}$ zusammengefasst, vollständig determiniert.

Zusätzlich zeichnen sich randomisierte Experimente im einfachsten Fall aus durch einen Vergleich einer eigens definierten und durchgeführten (Be-)Handlung (t) mit einer parallel definierten Auslassung der (Be-)Handlung (c) (Fisher & Wishart, 1930); dieses Charakteristikum wird zumeist als *Manipulation der unabhängigen Variablen* bezeichnet und ist gleichzusetzen mit einer durch den Forscher induzierten und kontrollierten Varianz der unabhängigen Varia-

blen, so dass diese im Zusammenhang mit einem randomisierten treatment-assignment *exogen*, d.h. die beobachtete Varianz der abhängigen Variablen Y_i erklärend, wird.

Dies gilt es auszuführen: Der Manipulation der unabhängigen Variablen und einem randomisierten treatment-assignment folgend wird die Fehlervarianz bei der Messung des kausalen Effektes minimiert, da *systematische Störeinflüsse*, die sowohl die Selektion von i zu t, als auch den potentiellen response-Wert $y_{i,t}$ bedingen könnten und basal in den Humanwissenschaften zu unterstellen sind, über die Experimentalbedingungen balanciert werden (Rosenbaum, 1995). Derartig wirkende Variablen werden zumeist als *konfundierende Stör- oder Kovariablen*, deren erhobenen Werte für i sich in dem Vektor $\vec{x}_i$ zusammenfassen lassen, klassifiziert; auch ungemessene Störvariablen können vorliegen, deren Werte für i folgend als Vektor $\vec{z}_i$ zusammengefasst werden. Die konfundierende Wirkung dieser Kovariablen auf den potentiellen response-Wert $y_{i,t}$ lässt sich allgemeingültig für i darstellen als:

$$y_{i,t} = y_{i,c} + \tau_i + h(\vec{x}_i, \vec{z}_i, \tau_i), \tag{3}$$

wobei $h(\vec{x}_i, \vec{z}_i, \tau_i)$ eine beliebige Funktion der Kovariablen und dem treatment-Effekt τ_i indiziert.

In Folge einer Randomisierung wird der Einfluss dieser Kovariablen auf den Selektionsindikator S_i ausgeschlossen, da die Selektion rein zufällig erfolgt; formal gilt unter Randomisierung in der Notation nach Dawid (1979): $S_i \perp\!\!\!\perp (\mathbf{X}, \mathbf{Z})$, wobei $\mathbf{X}_{N\times p}$ als Kovariablenmatrix über die $p-$gemessenen Kovariablen der N statistischen Einheiten definiert ist, $\mathbf{Z}$ über die entsprechenden nicht erhobenen Kovariablen. Der Unabhängigkeit der Kovariablen vom Selektionsindiktator folgt unmittelbar: $\Pr(S_i = 1 \mid \vec{x}_i, \vec{z}_i) = \Pr(S_i = 1)$, so dass die Wahrscheinlichkeit zu t selegiert zu werden für alle $\vec{x}_i$ und $\vec{z}_i$ identisch ist. In Konsequenz folgt eine identische Verteilung der Kovariablen über die Experimentalbedingungen, formal: $F(\mathbf{X} \mid S_i = 1) = F(\mathbf{X} \mid S_i = 0)$ sowie $F(\mathbf{Z} \mid S_i = 1) = F(\mathbf{Z} \mid S_i = 0)$, wobei $F(\cdot)$ die zugehörige Verteilungsfunktion indiziert (z.B., Imai et al., 2008; Rosenbaum, 1995). Die der randomisierten Selektion resultierenden Experimentalgruppen sind in ihren Kovariablen, die entsprechende Störeinflüsse darstellen können, direkt vergleichbar, d.h. *homogen*, und mögliche Effekte der Kovariablen auf den response sind über die Experimentalbedingungen hinweg balanciert. Als Modell für die response-Variable Y_i gilt entsprechend: $E(Y_i \mid S_i = s_i, \vec{x}_i, \vec{z}_i) = E(Y_i \mid S_i = s_i) + E(\epsilon_i \mid S_i = s_i)$ mit: $E(\epsilon_i \mid S_i = s_i) = 0,\ Cov(\epsilon_i, S_i) = 0$.

Observational Studies

Quasi-experimentelle Forschungsdesigns, auch *observational studies* (z.B., Cochran & Chambers, 1965; Rosenbaum, 1995), sind basal charakterisiert durch eine vollständige Kontrolle über die Ausprägungen der unabhängigen Variablen, d.h. durch eine eigenständige Definition der Menge K, und durch eine Selektion von i zu t, die nicht vom Forscher kontrolliert wird, so dass bei Datenerhebung die Selektion in eine der Bedingungen bereits vorliegt. Damit unterscheiden sich observational studies und randomisierte Experimente grundlegend in dem Modus, wie i jeweilig einer der Experimentalbedingungen zugewiesen wird. In Folge der fehlenden Randomisierung kann das Design in observational studies aus zwei Gründen heraus konfundiert sein:

Basal muss unterstellt werden, dass i sich auf Grund einer Kenntnis über seinen $y_{i,t}-$Wert selber t zuordnet - die potentiellen response-Werte wären damit nicht mehr unabhängig vom Selektionsindikator, so dass: $\Pr(S_i = 1 \mid y_{i,t}, y_{i,c}) \neq \Pr(S_i = 1)$. Eine derartige Konfundierung könnte dazu führen, dass sich alle N Einheiten vollständig t zuordnen und ein kausaler Rückschluss auf Grund fehlender Beobachtungen unterhalb von c nicht mehr möglich wäre, oder dass Einheiten, die von t profitieren, sich gezielt dem treatment zuordnen (z.B., Imbens & Rubin, 2012; Morgan & Winship, 2007; Wooldridge, 2003). Eine solche Konfundierung kann unterhalb eines randomisierten treatment-assignments stets ausgeschlossen werden, da $\Pr(S_i = 1 \mid y_{i,t}, y_{i,c}) = \Pr(S_i = 1)$ der Randomisierung folgend vorausgesetzt werden kann. Wie in folgenden Abschnitten gezeigt wird, hat eine derartige Konfundierung weitreichende Konsequenzen hinsichtlich der Identifikation interessierender kausaler Effekte.

Zumeist lässt sich diese Konfundierung ersetzen durch eine angenommene Konfundierung durch die Kovariablenwerte $\vec{x}_i$: In diesem Szenario erfolgt die Selektion von i zu t in Abhängigkeit von $\vec{x}_i$, jedoch kann bedingt an $\vec{x}_i$ ein randomisiertes treatment-assignment als Modell aufgestellt werden, so dass die Wahrscheinlichkeit, zu t selegiert zu werden, vollständig durch $\vec{x}_i$ erklärt und der Selektionsindikator konditional an $\vec{x}_i$ unabhängig von den potentiellen Werten ist. Als Selektionsmodell wird für observational studies $\Pr(S_i = 1 \mid \vec{x}_i, \vec{z}_i, y_{i,t}, y_{i,c}) = \Pr(S_i = 1 \mid \vec{x}_i)$ formuliert und damit das Problem einer Konfundierung durch die response-Werte aufgehoben. In Folge einer solchen Modellformulierung kann zunächst nicht mehr von einer Balance der Kovariablen im Vergleich beider Experimentalbedingungen ausgegangen werden, so dass diese Kovariablen als Störvariablen fungieren können, die damit sowohl die Selektionswahrschein-

lichkeit als auch die response-Werte beeinflussen. Jedoch lassen sich bedingt an $\vec{x}_i$ sämtliche, folgend aufzuführende Annahmen festhalten, die bei Randomisierung gelten und eine Identifikation kausaler Effekte zulassen. Der zweite Fall wird in vorliegender Arbeit von Interesse sein und an späterer Stelle als *strongly ignorable treatment-assignment* vollständig formalisiert.

1.5 Statistische Annahmen: Kausaler Effekt

Folgend sollen vorangestellte Ausführungen zur Messung eines *kausalen Effektes* durch die vollständige Vorstellung des *Rubin-Modell der kausalen Inferenz* ausgebaut werden und Lösungsmöglichkeiten für das Fundamentalproblem geschildert werden (z.B., Holland, 1986; Holland & Rubin, 1988; Imbens & Rubin, 2012).

Definitionen

- U: Finite Grundgesamtheit, verstanden als die gesamte Menge der statistischen Einheiten, die in räumlicher, sachlicher und zeitlicher Hinsicht für die Untersuchung von Interesse ist:

  $$U := \{u_i \mid i = 1, \ldots, N\},$$

  mit $i = 1, ..., N$ als Laufindex über die statistischen Einheiten.

- $K := \{c, t\}$: Menge der kausalen Agenten. Im einfachsten Design hat K zwei Elemente: a) ein treatment, $t \in K$, definiert als eine (Be-)Handlung, die einer Einheit i zukommen kann, b) eine Kontrollbedingung, $c \in K$, in der die (Be-)Handlung nicht appliziert wird. Im allgemeinen Fall indiziert $k \in K$ eine beliebige Experimentalbedingung, die in K aufgeführt wird.

- I_i : Selektionsindikator, der bei einer einfachen Stichprobenziehung eine Auswahl von i indiziert, wobei fortlaufend:

  $$I_i = \begin{cases} 1, & \text{wenn } i \text{ in Stichprobe} \\ 0 & \text{sonst.} \end{cases}$$

  Folgend wird $\Pr(I_i = 1) = \frac{1}{N}$, $\forall i \in U$, bei einer einfachen Stichprobenziehung eines i vorausgesetzt.

- τ_i : ist definiert als der individuelle kausale Effekt, der *potentiell* an einem aus U zufällig gezogenen i beobachtbar ist, wobei: $\tau_i := y_{i,t} - y_{i,c}$. Der Definition dieses Effektes nach wird folgend ein individueller additiver Effekt, der dem treatment folgt, unterstellt, so dass: $y_{i,t} = y_{i,c} + \tau_i$.

- Y_t : Zur Messung des potentiellen response-Wertes $y_{i,t}$ wird ein i aus U entnommen, unterhalb von t beobachtet und der potentielle response $y_{i,t}$ nach Applikation des treatments erhoben. Es ergibt sich in Folge der zufälligen Stichprobenziehung der Ergebnisraum Ω_i. Die Variable $Y_t(\omega_i)$ ordnet als Zuweisungsvorschrift allen Einheiten aus U den zugehörigen potentiellen response-Wert unterhalb von t zu, so dass die Realisierung, $Y_t(\omega_i) = y_{i,t}$, dem potentiellen $y_{i,t}$ für das gezogene i entspricht, welcher bei Exposition zu t beobachtet würde. Damit entspricht $Y_t(\omega_i)$ einer Abbildung auf dem Ergebnisraum, $Y_t(\omega_i) : \Omega_i \to \mathbb{R}$, und stellt eine Zufallsvariable dar. Der Einfachheit halber gilt folgend notationell $Y_t(\omega_i) := Y_t$.

  Ω_i lässt sich für $n > 1$ unmittelbar erweitern durch das kartesische Produkt

  $$\Omega_1 \times \Omega_2 \times ... \times \Omega_n,$$

  wobei n folgend die Stichprobengröße indiziert.

- Y_c ist selbig definiert wie Y_t, gibt an dieser Stelle als Realisierung jedoch den response $y_{i,c}$ von i bei Ziehung aus U an, der gemessen wird, wenn i in die Kontrollbedingung kommt.

- $\Delta := Y_t - Y_c$ indiziert die Differenz der Zufallsvariablen Y_t und Y_c, die einem wirkungsvollen t folgt. Die Realisierung $\Delta = \tau_i$ entspricht dem individuellen kausalen Effekt bei Beobachtung von i.

- S_i, mit $s_i \in \{0, 1\}$ als Bildmenge, ist definiert als der Selektionsindikator, der angibt, zu welcher Experimentalbedingung i faktisch zugewiesen wird. Wie bereits aufgeführt, erfolgt die Selektion von i in eine der Bedingungen im Idealfall randomisiert und es gilt folgend: wird $S_i = 1$ realisiert, wird i zu t selegiert; wird $S_i = 0$ realisiert, wird i der Kontrollbedingung ausgesetzt. Die Realisierung $S_i = s_i$ entscheidet stets darüber, welcher der beiden potentiellen Werte $y_{i,t}$ und $y_{i,c}$ an i beobachtet werden kann, denn derjenige Wert der Bedingung, zu der i nach der Selektion nicht zugewiesen wird, bleibt *kontrafaktisch*, d.h. nach der Selektion nicht mehr beobachtbar, obwohl er weiterhin messbar wäre, wenn

es die Selektion nicht gegeben hätte. Gegeben diesem *Fundamentalproblem der kausalen Inferenz* lässt sich der faktisch beobachtete response von i, y_i, darstellen als:

$$y_i = s_i \cdot y_{i,t} + (1 - s_i) \cdot y_{i,c}, \quad \text{für } s_i \in \{0, 1\}$$

- N_t : Menge der Einheiten, die zu t selegiert werden - mit $s_i \in \{0, 1\}$ gilt: $N_t = \sum_{i=1}^{N} S_i$; entsprechend: $N_c = \sum_{i=1}^{N}(1 - S_i)$ und $N = N_t + N_c$

- $\mathbf{X}_{N \times p}$: Matrix der erhobenen Kovariablenwerte über $p-$Kovariablen; der $i-$te Zeilenvektor $\vec{x}_i$ entspricht dem Vektor der Kovariablenwerte der Einheit i

- $\mathbf{Z}_{N \times p}$: Matrix der nicht erhobenen Kovariablenwerte über $p-$Kovariablen.

Die mögliche Messung des kausalen Effektes τ_i beruht auf einer, aus dem Kontext der Untersuchung heraus definierten Grundgesamtheit U, aus der i gezogen und *potentiell* zwei Experimentalbedingungen zugeführt werden könnte; dem Modell nach kann unterhalb jeder der Bedingungen ein *fester* Messwert beobachtet werden $(y_{i,t}, y_{i,c})$. Eine Beobachtung des Effektes τ_i würde unmittelbar Aufschluss über die individuelle Wirkung von t geben, so dass eine fortlaufende Beobachtung weiterer statistischer Einheiten Auskunft über den in U zu Grunde liegenden Effekt Δ geben würde, der in allgemeingültiger Hinsicht von Interesse ist.

In praxi ist es jedoch nicht möglich, das Wertepaar $(y_{i,t}, y_{i,c})$ an i zeitgleich, oder wohl möglich zeitversetzt, zu beobachten; wird i t ausgesetzt, so lässt sich in den Humanwissenschaften unterstellen, wird sie sich verändern und eine anschließende Messung von $y_{i,c}$ wird bedingt durch die vorherige Exposition zu t (z.B., Holland, 1986; Morgan, 2001). Entsprechend muss i in eine Versuchsbedingung selegiert werden. Holland und Rubin (1988, S.206) weisen in diesem Zusammenhang auf:

> „A question that immediately arises is whether or not it is *ever possible* to expose a unit to more than one treatment [...]. One can argue that this is never possible in principle, because once a unit has been exposed to a treatment, the unit is different from what it was before.“

In Folge der notwendigen Selektion bleibt der Effekt τ_i stets unbeobachtbar und es liegen damit ausgehend von i keine Informationen über den generellen Effekt Δ vor. Wie fortlaufend zu zeigen ist, lässt sich dieses bestehende Fundamentalproblem unter Gültigkeit zu spezifizierender

Restriktionen, die an den Selektionsindikator S_i gestellt werden, dahin gehend lösen, als dass eine kausale Inferenz nicht alleinig auf der Beobachtung eines i sondern auf der Beobachtung von Gruppen statistischer Einheiten, die anteilsmäßig zu t oder c selegiert werden, beruht.

Der Selektionsindikator S_i und Restriktionen

Es gilt folgend, drei basale Restriktionen an den Selektionsindikator S_i aufzustellen, unterhalb der Gültigkeit dieser das *Fundamentalproblem der kausalen Inferenz* lösbar wird.

1. *probabilistisches treatment-assignment*: Es wird folgend mit einer faktischen Selektion von i in eine der Bedingungen vorausgesetzt, dass $S_i = s_i$ in jedem Falle hätte anders ausfallen können; diese Restriktion findet sich als *probabilistisches treatment-assignment* in der Literatur (z.B., Holland, 1986; Imbens & Rubin, 2012) und formuliert damit statistisch die *potential exposability*-Annahme aus. Die potential exposability lässt sich verdichten zu der Annahme, dass eine Kausalaussage nur getroffen werden kann, wenn die Ausprägungen der unabhängigen Variablen manipulierbar sind, da nur an entsprechender Stelle das *ceteris paribus*-Prinzip angewandt werden kann. Die Forderung der potential exposability schließt damit Charakterstika von Personen als kausale Agenten aus; kausale Inferenzen können nur in Verbindung mit treatments, zu denen i jeweilig eine *positive Chance* auf Exposition hat, getätigt werden. Holland (1986, S. 946) hebt hervor:

> „For causal inference, it is critical that each unit be *potentially exposable* to any of the causes. (...) In a controlled study, S is constructed by the experimenter. In an uncontrolled study, S is determined to some extent by factors beyond the experimenter's control. In either case, the critical feature of the notion of cause in this model is that the value of $S(u)$ for each unit *could have been different*."

Ein probabilistisches treatment-assignment ist damit abhängig von einer zutreffenden potential exposability-Annahme: Nur, wenn i beiden Experimentalbedingungen potentiell zugewiesen werden kann, kann diese Selektion zufällig erfolgen und ist nicht determiniert, wie in Folge von Charakteristika. Formal gilt ein probabilistisches treatment-assignment erfüllt, wenn:

$$0 < \Pr(S_i = 1) < 1, \quad \forall\, i \in U.$$

Es ist unmittelbar ersichtlich, dass in randomisierten Experimenten diese Annahme in Folge eines exemplarischen Münzwurfes, der die Selektion von i determiniert, erfüllt wird, solange

die Selektionswahrscheinlichkeit stets im Wertebereich der positiven Wahrscheinlichkeit liegt. Wie an späterer Stelle aufgewiesen wird, lässt sich in observational studies ebenfalls ein probabilistisches treatment-assignment modellieren, wobei die probabilistische Selektion bedingt an $\vec{x}_i$ erfolgt, so dass als Restriktion für observational studies folgend

$$0 < \Pr(S_i = 1 \mid \vec{x}_i) < 1, \quad \text{für alle möglichen } \vec{x}_i.$$

aufgestellt wird.

2. *unconfounded treatment-assignment*: Wie in den nächsten Abschnitten formalisiert wird, setzt die Lösung des Fundamentalproblems stets ein *ignorable* bzw. *unkonfundiertes treatment-assignment* voraus (z.B., Holland, 1986; Imbens & Rubin, 2012), unterhalb dessen die potentiellen response-Variablen stochastisch unabhängig vom Selektionsindikator sind, formal wird in Dawid's Notation gefordert:

$$(Y_t, Y_c) \perp\!\!\!\perp S_i.$$

Es gilt folglich für die Selektionswahrscheinlichkeit:

$$\Pr(S_i = 1 \mid y_{i,t}, y_{i,c}) = \Pr(S_i = 1), \quad 0 < \Pr(S_i = 1) < 1, \forall i \in U.$$

Damit tragen die potentiellen response-Werte $y_{i,t}$ und $y_{i,c}$ keinerlei Informationen über die Selektion von i zu t. Ein unkonfundiertes treatment-assignment ist stets bei Randomisierung erfüllt, da i dem Zufall nach zu t oder c zugewiesen wird - sie sich demnach nicht durch Kenntnis über die Wirkung von t selber in diese Bedingung selegiert. In observational studies, innerhalb dessen sich die Einheiten de facto selber selegieren, kann ein *unkonfundiertes treatment-assignment* nicht bedingungslos vorausgesetzt werden, wie Morgan und Winship (2007, S.79) betonen:

> „(...) some individuals are thought to enter the programs based on anticipation of the treatment effect itself (...).“

Die Lösung einer stets zu unterstellenden Konfundierung durch die response-Werte in observational studies beruht auf der Modellvorstellung einer probabilistischen Selektion von i zu t, die vollständig durch die beobachteten Kovariablenwerte $\vec{x}_i$ erklärt wird. Dem folgend kann eine an $\vec{x}_i$ bedingte Unabhängigkeit der response-Variablen vom Selektionsindikator unterstellen werden, so dass in Dawid's Notation:

$$(Y_t, Y_c) \perp\!\!\!\perp S_i \mid \vec{x}_i.$$

Ein solches treatment-assignment wird folgend als *strongly ignorable treatment-assignment* (z.B., Rosenbaum, 1984b; Rosenbaum & Rubin, 1983) bezeichnet; z.T. findet sich auch die Bezeichnung der *selection on observables* (z.B., Heckman, 2005; Imai et al., 2008). Diese Annahme führt dazu, dass in observational studies das treatment-assignment als an $\vec{x}_i$ bedingt unkonfundiert modelliert werden kann, so dass folglich:

$$\Pr(S_i = 1 \mid \vec{x}_i, y_{i,t}, y_{i,c}) = \Pr(S_i = 1 \mid \vec{x}_i), \quad 0 < \Pr(S_i = 1 \mid \vec{x}_i) < 1, \text{ für alle } \vec{x}_i.$$

Die Relevanz der Unkonfundierung soll an dieser Stelle bereits hervorgehoben werden und fortlaufend formalisiert werden: In Folge eines randomisierten treatment-assignments gilt stets, dass alle gemessenen und ungemessenen Variablen, die in einer stochastischen Beziehung zu den potentiellen response-Variablen stehen könnten, unabhängig von S_i sind: das realisierte $S_i = s_i$ ist stets ein Zufallsprodukt, beispielsweise in Folge eines Münzwurfs, und zwar unabhängig von den Kovariablenwerten $\vec{x}_i$ und $\vec{z}_i$ und den potentiellen response-Werten $y_{i,t}$ und $y_{i,c}$. Rosenbaum (2010, S.27) verdeutlicht entsprechend:

> „(...) the coin is fair not just in coming up heads half the time, independently (...), but more importantly the coin knows nothing about the individual and is impartial in its treatment assignments.“

Für eine randomisierte Selektion der N Einheiten folgt damit, dass die der Selektion resultierenden Gruppen, wobei N_t-Einheiten dem treatment und N_c-Einheiten der Kontrollbedingung zugewiesen werden, gegeben einem unkonfundierten treatment-assignment vollständig homogen in allen Variablen sind; der Unabhängigkeit des Selektionsindikators von den potentiellen response-Variablen folgt:

$$E(Y_i \mid S_i = 1) = E(Y_t \mid S_i = 1) = E(Y_t \mid S_i = 0) \text{ sowie:}$$
$$E(Y_i \mid S_i = 0) = E(Y_c \mid S_i = 0) = E(Y_c \mid S_i = 1),$$

so dass sich die der Selektion resultierenden Gruppen im ersten Moment der Verteilung in den potentiellen response-Werten gleichen und der identifizierbare Erwartungswert $E(Y_t \mid S_i = 1)$ als Informationsquelle für den kontrafaktischen Erwartungswert $E(Y_t \mid S_i = 0)$ genutzt werden kann; trotz vorliegender Selektion liefern beide Gruppen alle benötigten Informationen für den Rückschluss auf die Wirkung von t. Da gegeben einem randomisierten treatment-assignment zusätzlich

$$(\mathbf{X}, \mathbf{Z}) \perp\!\!\!\perp S_i$$

gilt, folgt der Randomisierung zusätzlich:

$$F(\mathbf{X} \mid S_i = 1) = F(\mathbf{X} \mid S_i = 0), \text{ sowie: } F(\mathbf{Z} \mid S_i = 1) = F(\mathbf{Z} \mid S_i = 0).$$

Damit sind die einzigen messbaren Unterschiede zwischen den Einheiten nach einer randomisierten Selektion die Zugehörigkeiten zu einer der Experimentalbedingungen sowie unter Wirkung des treatments die Varianz in der beobachteten Variablen Y_i. Damit kann diejenige Gruppe unterhalb von t die Informationen über die Wirkung des treatments liefern, die der Gruppe unterhalb von c vorenthalten blieb - und vice versa. Wie zu zeigen ist, setzt hier die Lösung des Fundamentalproblems an.

3. *individualisitic treatment-assignment*: Zusätzlich wird folgend ein *individualistisches treatment-assignment* vorausgesetzt, unterhalb dessen die Zuordnungswahrscheinlichkeit von i unabhängig von allen anderen N Einheiten ist; im vollständig randomisierten Fall lässt sich bei Gültigkeit der Unkonfundierung als Selektionsmechanismus festhalten:

$$\begin{aligned}
&\Pr(S_1 = s_1, S_2 = s_2, \ldots, S_N = s_N \mid y_{1,t}, y_{1,c}, \vec{x}_1, \vec{z}_1, y_{2,t}, y_{2,c}, \vec{x}_2, \vec{z}_2, \ldots, y_{N,t}, y_{N,c}, \vec{x}_N, \vec{z}_N) \\
=&\Pr(S_1 = s_1, S_2 = s_2, \ldots, S_N = s_N) \\
=&\prod_{i=1}^{N} \Pr(S_i = s_i) = \prod_{i=1}^{N} \pi^{s_i}(1-\pi)^{(1-s_i)},
\end{aligned}$$

wobei: $\Pr(S_i = 1) = \pi, \pi \in (0,1)$ als Indizierung der festen Selektionswahrscheinlichkeit. Damit lässt sich *im einfachsten Fall* für ein randomisiertes treatment-assignment als Verteilungsmodell ein Bernoulli-Versuch aufstellen mit: $S_i \sim B(1,\pi),\ 0 < \pi < 1$.

Auch in observational studies wird folgend ein individualistic treatment-assignment vorausgesetzt. Gegeben der vorangestellten Modellannahmen lässt sich als Modell in observational studies festhalten:

$$\begin{aligned}
&\Pr(S_1 = s_1, S_2 = s_2, \ldots, S_N = s_N \mid y_{1,t}, y_{1,c}, \vec{x}_1, \vec{z}_1, y_{2,t}, y_{2,c}, \vec{x}_2, \vec{z}_2, \ldots, y_{N,t}, y_{N,c}, \vec{x}_N, \vec{z}_N) \\
=&\Pr(S_1 = s_1, S_2 = s_2, \ldots, S_N = s_N \mid \vec{x}_1, \vec{x}_2, \ldots, \vec{x}_N) \\
=&\prod_{i=1}^{N} \Pr(S_i = s_i \mid \vec{x}_i)
\end{aligned}$$

Dieses Modell formuliert damit die Annahme des *strongly ignorable treatment-assignment* aus und wird an späterer Stelle ausgeführt.

Stable Unit Treatment Value Assumption (SUTVA)

Den vorherigen Ausführungen, insb. (2), zugrunde liegt die *stable unit treatment value assumption*, kurz *SUTVA*, die folgend expliziert werden soll, da sie nebst Formulierung der potentiellen Werte die Grundlage des Rubin-Modells bildet und unmittelbar eine entscheidende Restriktion bei der Beobachtung von N Einheiten, die folgend als Lösung des Fundamentalproblems eingeführt wird, formuliert (z.B., Holland & Rubin, 1988; Rubin, 1974). Der *SUTVA* nach ist der *beobachtete response* y_i gleichzusetzen mit dem potentiellen $y_{i,k}-$Wert, das demjenigen $k \in K$ resultiert, zu dem i selegiert wird. Somit besitzt i für alle definierten Experimentalbedingungen jeweilig einen potentiellen response-Wert, der der Annahme nach ausschließlich durch die jeweilige Handlung unterhalb des k evoziert wird und bei entsprechender Selektion beobachtet wird. Rubin (2007, S.773) führt aus:

> „The most straightforward assumption to make is the 'stable unit treatment value assumption'(SUTVA) under which the potential outcome for the ith unit just depend on the treatment the ith unit received.“

Damit impliziert die *SUTVA* zum Einen, dass die Explikation der Experimentalbedingungen für alle N Einheiten identisch ist (*no hidden variations in treatment*, Imbens und Rubin (2012)), zum Anderen, dass die Beobachtung des potentiellen Wertes, $y_{i,k}$, nur auf den individuellen Selektionsstatus, $S_i = s_i$ zurückzuführen ist, jedoch nicht auf ein vollständig realisiertes treatment-assignment, $\vec{S} = \vec{s}$, das die Selektionen aller N Einheiten in dem Zufallsvektors $\vec{S}_{N\times 1}$ zusammenfasst (z.B., Imbens & Rubin, 1997, 2012; Little & Rubin, 2000). Damit werden die potentiellen Werte als *stabil* aufgefasst, d.h., dass sie sich ausschließlich aus der Durchführung der Experimentalbedingungen ergeben und sich nicht in Folge eines realisiertes treatment-assignment anderer statistischen Einheiten, wie auch Imbens und Rubin (2012, S.12) hervorheben:

> „The point is that SUTVA implies that the potential outcomes for each unit and each treatment are well-defined functions (possibly with stochastic images) of the unit index and the treatment.“

Morgan und Winship (2007, S. 38) führen ein Beispiel auf, welches eine mögliche Verletzung der *SUTVA* verdeutlichen soll: Angenommen, es liegen $N = 3$ Einheiten vor und das treatment-assignment erfolgt randomisiert, so könnten mit $N_t = 1$ und $\Pr(S_i = 1) = \frac{1}{3}$ folgende $\binom{3}{1} = 3$

treatment-assignments in Form der Vektoren $\vec{S} = \vec{s}$ mit den entsprechenden Selektionsindikationen realisiert, und unabhängig davon, ob $\vec{s}_1, \vec{s}_2$ oder $\vec{s}_3$ realisiert würde, folgende potentiellen response-Werte beobachtet werden:

$$\vec{s}_1 = \begin{pmatrix} 1 \\ 0 \\ 0 \end{pmatrix} \quad \vec{s}_2 = \begin{pmatrix} 0 \\ 1 \\ 0 \end{pmatrix} \quad \vec{s}_3 = \begin{pmatrix} 0 \\ 0 \\ 1 \end{pmatrix}$$

i	$y_{i,t}$	$y_{i,c}$	τ_i
1	3	0	3
2	3	0	3
3	3	0	3

Eine Verletzung der *SUTVA* würde einhergehen mit einem zum oben aufgeführten alternierenden Selektionsmuster, bei dem zwei Einheiten das treatment $\left(\Pr(S_i = 1) = \frac{2}{3}\right)$ erhalten und sich die potentiellen response-Werte im Verhältnis zum vorherigen Szenario verändern:

$$\vec{s}_1 = \begin{pmatrix} 1 \\ 1 \\ 0 \end{pmatrix} \quad \vec{s}_2 = \begin{pmatrix} 1 \\ 0 \\ 1 \end{pmatrix} \quad \vec{s}_3 = \begin{pmatrix} 0 \\ 1 \\ 1 \end{pmatrix}$$

i	$y_{i,t}$	$y_{i,c}$	τ_i
1	2	0	2
2	2	0	2
3	2	0	2

Das Beispiel weist die Verletzung der *SUTVA* eindeutig auf: Nicht von dem individuellen Selektionsstatus $S_i = s_i$, sondern vom jeweiligen treatment-assignment $\vec{S}$ ist der potentielle response-Wert sowie individuelle kausale Effekt τ_i abhängig.

Insbesondere in der Ökonometrie finden sich Kritiker der *SUTVA*, die grundsätzlich eine *Interferenz* der Einheiten, die sich je nach treatment-assignment $\vec{S}$ auf die potentiellen response-Werte auswirkt, unterstellen und stattdessen einen Unterschied zwischen einem *general equilibrium Effekt* und einem *partial equilibrium Effekt*, der dem individuellen kausalen Effekt bei Konstanthaltung der Effekte und Interaktionen der anderen Einheiten angibt, vornehmen (Imbens & Rubin, 2012).

Average Causal Effect als statistische Lösung des Fundamentalproblems

Bereits Neyman (1932, übersetzt veröffentlicht in: Splawa-Neyman, Dabrowska & Speed, 1990) stand vor dem fundamentalen missing-data-Problem (vgl. (2)), das sich an dieser Stelle einstellte, als die Wirkung verschiedener treatments auf den Ernteertrag eines Feldes evaluiert werden sollte: De facto bleibt der individuelle Effekt τ_i, der die Wirkung von t unmittelbar quantifiziert, unbeobachtbar und i liefert damit keinerlei Informationen über den generell in U zu Grunde liegenden Effekt $\Delta := Y_t - Y_c$. Neyman (1932) führt auf, dass trotz der fehlenden Information von i die Lösung des Fundamentalproblems auf den Zufallsvariablen Y_t und Y_c beruht, die ei-

ne entsprechende Verteilung der potentiellen response-Werten in U indizieren; Lunceford und Davidian (2004, S.3) führen aus:

> „The distributions of Y_0 and Y_1 may be thought of as representing the hypothetical distributions of response for the population of individuals were all individuals to receive control or be treated, respectively, so the means of these distributions correspond to the mean response if all individuals were to receive each treatment.“

In Folge eines wirkungsvollen treatments sollten sich die Verteilungen dieser Zufallsvariablen, $F(Y_t)$ und $F(Y_c)$, sowie in Konsequenz auch die zugehörigen Erwartungswerte als erstes Moment, $E(Y_t)$ und $E(Y_c)$, unterscheiden. Die Differenz der Erwartungswerte, $E(Y_t) - E(Y_c)$, ist in der Literatur üblicherweise als der *average causal effect* (ACE), als *average treatment effect* oder als τ definiert und war bereits für Neyman (1932) als Lösung des Fundamentalproblems von Interesse: Der ACE wird im Fall eines finiten U interpretiert als der durchschnittliche kausale Effekt, der sich quantifizieren ließe, wenn alle N Einheiten unterhalb beider Experimentalbedingungen beobachtet und die resultierenden τ_i gemittelt werden (z.B., Guo & Fraser, 2010; Holland, 1986; Lunceford & Davidian, 2004; Morgan & Winship, 2007; Splawa-Neyman et al., 1990). Entsprechend der Bedeutung als Erwartungswert lässt sich der ACE inhaltlich explizieren durch:

> „The average causal effect is also equal to the expected value of the what-if difference (...) for a randomly selected (...) from the population.“ (Morgan & Winship, 2007, S.37)

Formal ist der ACE mit $\Pr(I_i = 1) = \frac{1}{N}$ definiert als:

$$\tau := E(\Delta) = E(Y_t) - E(Y_c) = \frac{1}{N} \cdot \sum_{i=1}^{N} y_{i,t} - \frac{1}{N} \cdot \sum_{i=1}^{N} y_{i,c} = \frac{1}{N} \cdot \sum_{i=1}^{N} \tau_i \tag{4}$$

Grundsätzlich bleibt das Fundamentalproblem der kausalen Inferenz auch bei der Definition des ACE bestehen, so dass der ACE zunächst keine Lösungsmöglichkeit darstellt: Der ACE ist ohne weitere Annahmen nicht identifizierbar, da die Quantifizierung des ACE auf einer, faktisch nicht möglichen, simultanen Exposition aller Einheiten zu beiden Experimentalbedingungen basiert. In praxi ist jedoch eine anteilsmäßige Selektion der N Einheiten in eine der Bedingungen möglich, so dass die bedingten Erwartungswerte $E(Y_i \mid S_i = 1) = E(Y_t \mid S_i = 1)$ und $E(Y_i \mid S_i = 0) = E(Y_c \mid S_i = 0)$ der Selektion entsprechend identifiziert werden können.

Wie folgend zu zeigen ist, folgt einem randomisierten treatment-assignment eine Identifikation des ACE anhand dieser bedingten Erwartungswerte; somit beruht die Lösung des Fundamentalproblems, wie von Neyman (1932) bereits aufgewiesen, auf einer anteilsmäßigen Beobachtung von N_t Einheiten unterhalb von t und N_c Einheiten unterhalb von c. Der Randomisierung folgend lässt sich mit dem ACE ein Parameter identifizieren, der einerseits einen Schätzwert für den unbeobachtbaren Effekt τ_i (Bedeutung des Erwartungswertes) darstellt, andererseits in Stichproben erwartungstreu und konsistent schätzbar ist, solange die an den Selektionsindikator S_i aufgestellten Restriktionen gültig sind.

Unit Homogenity, Randomisierung und Additiver Effekt

Rubin (1974) stand in direkter Tradition von Neyman und schlug zwei Lösungsmöglichkeiten für das Fundamentalproblem vor, die in dem Artikel anhand eines *two-trials*-Experimentes, d.h. bei $N = 2$, exemplifiziert werden: Analog zu Neyman wird basal die Lösung des Fundamentalproblems in einer anteilsmäßigen Beobachtung der N statistischen Einheiten in beiden Bedingungen gesehen, wobei der erste Lösungsvorschlag eine *unit homogenity* der Einheiten formuliert und Annahmen über den Modus der Selektion ausspart, während der andere Lösungsvorschlag den ACE als zentralen kausalen Parameter fokussiert und die Lösung des Fundamentalproblems in einem randomisierten treatment-assignment sieht.

Die *unit homogenity* formuliert die vereinfachende Annahme, dass alle N Einheiten denselben potentiellen response-Wert unterhalb t besitzen und auch unterhalb von c identische potentielle response-Werte vorliegen, so dass:

$$y_{1,t} = y_{2,t} = \ldots = y_{N,t} \text{ sowie } y_{1,c} = y_{2,c} = \ldots = y_{N,c}.$$

In Konsequenz wird damit ein konstanter Effekt $\tau_i = y_{i,t} - y_{i,c}$, $\forall i \in U$, formuliert, so dass: $\tau_1 = \tau_2 = \ldots = \tau_N$. Für das beispielhafte two-trial-Experiment mit $N = 2$ folgt unter dieser Annahme: $y_{1,t} = y_{2,t}$ sowie $y_{1,c} = y_{2,c}$. Demnach würde es genügen, wenn $i = 1$ zu t und $i = 2$ zu c selegiert wird - die beobachtete Differenz $y_{1,t} - y_{2,c}$ entspräche an dieser Stelle dem bisher unbeobachtbaren Effekt τ_i, da:

$$y_{1,t} - y_{2,c} = y_{2,t} - y_{1,c} = y_{1,t} - y_{1,c} = y_{2,t} - y_{2,c} = \tau_1 = \tau_2$$

Damit würde $i = 2$, wenn in c beobachtet, Informationen über den kontrafaktischen $y_{1,c}-$Wert liefern - und vice versa. Unterhalb dieser Annahme ist ersichtlich, dass $i = 2$ genauso zu t

selegiert werden kann und $i = 1$ zu c - der individuelle Effekt τ_i ist mit der Selektion beider Einheiten in eine der Bedingungen beobachtbar.

Die strigente unit homogenity-Annahme, wie hier aufgeführt, wird explizit in den Humanwissenschaften als haltlos angesehen (z.B., Höfler, 2005; Holland, 1986; Imai et al., 2008; Rubin, 1974), da sie jegliche Varianz in der Verteilung der potentiellen Werten, Y_t und Y_c, sowie mögliche Unterschiede zwischen den individuellen kausalen Effekten, τ_i, negiert und zusätzlich vollständig reliable Skalen bei der Erfassung von Y_t und Y_c voraussetzt.

Statt der unit homogenity, die zumeist nur in den Naturwissenschaften plausibel erscheint, wird in den Human- und Sozialwissenschaften der vorangestellte *average treatment effect* als Lösungsmöglichkeit für das bestehende Fundamentalproblem fokussiert. Es bleibt zu zeigen, dass die Lösung dieses Problems auf einem randomisierten treatment-assignment der N Einheiten in eine der beiden Bedingungen basiert. Zunächst soll der Zusammenhang von ACE und Randomisierung anhand des exemplarischen two-trial-Experiment dargestellt werden, bevor er allgemein gültig verallgemeinert wird:

Mit $N = 2$ steht für ein randomisiertes treatment-assignment die Selektionswahrscheinlichkeit $\Pr(S_i = 1) = \frac{1}{2}$ fest, so dass folglich die Beobachtung von $i = 1$ in t gleichwahrscheinlich wie die Beobachtung von $i = 2$ in t ist. Dem individuellen Effekt τ_i angelehnt lässt sich die Zufallsvariable $S_1 \cdot (y_{1,t} - y_{2,c}) + (1 - S_1) \cdot (y_{2,t} - y_{1,c})$ definieren, die als Realisierungsmöglichkeiten eine von der Selektion der Einheit $i = 1$ abhängige Beobachtung einer der Differenzen der potentiellen response-Werte besitzt (z.B., Imbens & Rubin, 2012; Rubin, 1974), d.h.:

$$S_1 \cdot (y_{1,t} - y_{2,c}) + (1 - S_1) \cdot (y_{2,t} - y_{1,c}) = \begin{cases} (y_{1,t} - y_{2,c}), & \text{wenn } S_1 = 1 \\ (y_{2,t} - y_{1,c}), & \text{wenn } S_1 = 0. \end{cases}$$

Der Zusammenhang zum ACE, wobei hier $\text{ACE} = \frac{1}{2}(y_{1,t} - y_{1,c}) + \frac{1}{2}(y_{2,t} - y_{2,c})$, lässt sich unmittelbar herstellen, da diese Zufallsvariable unterhalb eines randomisierten treatment-assignments eine erwartungstreue Schätzstatistik für den ACE darstellt (z.B., Imai et al., 2008; Imbens & Rubin, 2012; Rubin, 1974), es gilt:

$$\begin{aligned} E\left(S_1 \cdot (y_{1,t} - y_{2,c}) + (1 - S_1) \cdot (y_{2,t} - y_{1,c})\right) &= \frac{1}{2}(y_{1,t} - y_{2,c}) + \frac{1}{2}(y_{2,t} - y_{1,c}) \\ &= \frac{1}{2}\left((y_{1,t} - y_{1,c}) + (y_{2,t} - y_{2,c})\right) \\ &= \text{ACE} \end{aligned}$$

Somit könnten Replikationen des Experimentes sowie der Gebrauch erwartungstreuer Schätzstatistiken Aufschluss über den ACE geben, wie es Neyman (1932) entsprechend als Lösungsmöglichkeit sah. In weiterführender Hinsicht lässt sich für ein finites N festhalten:

> „The Neyman-Rubin counterfactual framework holds that a researcher can estimate the counterfactual by examining the average outcome of the treatment participants and the average outcome of the nontreatment participants in the population. That is, the researcher can assess the counterfactual by evaluting the difference in mean outcomes between the two groups or 'averaging out' the outcome values of all individuals in the same condition.“ (Guo & Fraser, 2010, S.25)

Damit erscheint es in einem finiten N naheliegend, als Schätzstatistik für den ACE die nach der Selektion resultierende Differenz der gruppenspezifischen Mittel zu wählen, d.h.:

$$\hat{\text{ACE}} = \left(\frac{1}{N_t}\sum_{i:S_i=1} y_i\right) - \left(\frac{1}{N_c}\sum_{i:S_i=0} y_i\right) = \frac{1}{N}\sum_{i=1}^{N}\left(\frac{S_i \cdot y_{i,t}}{\pi} - \frac{(1-S_i)\cdot y_{i,c}}{1-\pi}\right) \tag{5}$$

Die Erwartungstreue lässt sich unter der Annahme fester potentieller response-Werte (SUTVA) und einer festen Selektionswahrscheinlichkeit, $\Pr(S_i = 1) = \pi$, unmittelbar aufweisen:

$$\begin{aligned} E(\hat{\text{ACE}}) &= \frac{1}{N}\sum_{i=1}^{N}\left(\frac{E(S_i)\cdot y_{i,t}}{\pi} - \frac{E((1-S_i))\cdot y_{i,c}}{1-\pi}\right) \\ &= \frac{1}{N}\sum_{i=1}^{N}\left(\frac{\pi \cdot y_{i,t}}{\pi} - \frac{(1-\pi)\cdot y_{i,c}}{1-\pi}\right) \\ &= \frac{1}{N}\sum_{i=1}^{N}(y_{i,t} - y_{i,c}) = \text{ACE} \end{aligned}$$

Damit folgt, dass die Lösung des Fundamentalproblems auf einer anteilsmäßigen, randomisierten Selektion der $N-$Einheiten in beide Bedingungen beruht und Schätzstatistiken formuliert werden können, die den ACE erwartungstreu schätzen. Die vorherigen Ausführungen werden im Folgenden in einen Zusammenhang mit einfachen Zufallsstichproben gebracht, die in praxi dazu genutzt werden, den ACE zu schätzen.

1.6 Average Causal Effect (ACE)

Im Folgenden werden die Restriktionen, die an den Selektionsindiktor S_i gestellt werden (vgl. S. 20ff.), zusammengebracht mit dem ACE, denn unter Gültigkeit der aufgewiesenen Annahmen

lässt sich zeigen, dass 1. der ACE anhand beobachtbarer Daten als messbare Größe identifiziert ist und dass 2. die Differenz der arithmetischen Mittel unterhalb der Experimentalbedingungen t und c als Schätzstatistiken, d.h. $g_t(Y_1, \ldots, Y_n) = \bar{Y}_t$ sowie $g_c(Y_1, \ldots, Y_n) = \bar{Y}_c$, in realisierten Stichproben erwartungstreu und konsistent den ACE schätzen. Holland (1986, S.947) hebt diesen Punkt hervor:

> „The important point is that the statistical solution replaces the impossible-to-observe causal effect of t on a specific unit with the possible-to-estimate *average* causal effect of t over a population of units.“

1.6.1 Schätzung des ACE

Es lässt sich bei der Schätzung des ACE in praxi ein Unterschied zwischen randomisierten Experimenten und observational studies treffen: in beiden Fällen wird aus U eine *einfache Zufallsstichprobe* vom Umfang n entnommen, so dass: $\Pr(I_i = 1) = \frac{n}{N}$ vorausgesetzt werden kann. In randomisierten Experimenten werden der Stichprobenziehung folgend mit einer festen Wahrscheinlichkeit $\Pr(S_i = 1) = \pi$ anteilsmäßig Einheiten t zugewiesen, folglich resultiert eine Gruppe, die dem treatment zugewiesen ist mit $n_t := \sum_i I(I_i = 1, S_i = 1)$, während ein anderer Anteil in die Kontrollbedingung selegiert wird ($n_c := n - n_t$); in praxi gilt im einfachsten Fall zumeist: $\pi = \frac{n_t}{n}$ (Imbens & Rubin, 2012), wobei mögliche Abweichungen von dieser Annahme später aufgewiesen werden. In observational studies liegen die Gruppen n_t und n_c bereits bei Datenerhebung vor, so dass die Selektion entsprechend vollzogen ist.

Für beide Fälle gilt: In Abhängigkeit von der Bedingung, in die i selegiert wird, entspricht unter Gültigkeit der SUTVA der beobachtete response-Wert y_i dem jeweiligen potentiellen Wert, der durch die Experimentalbedingung evoziert wird (vgl. (2)). Damit liegen in der Stichprobe dem treatment-assignment folgend Realisierungen der Variablen Y_t und Y_c vor. Als Schätzstatistik für den ACE wird in praxi zumeist die Stichprobenfunktion

$$\begin{aligned} \bar{Y}_t - \bar{Y}_c &= \frac{1}{n_t} \sum_{i:i \in \{I_i=1, S_i=1\}} y_i - \frac{1}{n_c} \sum_{i:i \in \{I_i=1, S_i=0\}} y_i \\ &= \frac{1}{n_t} \sum_{i:i \in \{I_i=1, S_i=1\}} y_{i,t} - \frac{1}{n_c} \sum_{i:i \in \{I_i=1, S_i=0\}} y_{i,c} \end{aligned} \tag{6}$$

genutzt, wobei sich diese „naïve Schätzung“ (Morgan & Winship, 2007, S. 44) des ACE unmittelbar mit der SUTVA in Verbindung bringen lässt, da unter ihrer Gültigkeit basal *erhofft*

wird, dass diese Statistik erwartungstreu sowie konsistent den ACE schätzt, d.h.:

$$E\left(\bar{Y}_t - \bar{Y}_c\right) = E(Y_t) - E(Y_c)$$

$$\bar{Y}_t - \bar{Y}_c \xrightarrow{f.s.} E(Y_t) - E(Y_c)$$

Diese Annahme ist jedoch nicht bedingungslos haltbar, da in Folge der notwendigen Selektion die Schätzfunktionen $\bar{Y}_t$ und $\bar{Y}_c$ zunächst einmal erwartungstreue und konsistente Schätzer für diejenigen Erwartungswerte darstellen, die an dem Ereignis $S_i = s_i$ bedingt sind (vgl. Laufindex in (6)). Es gilt damit zunächst grundsätzlich:

$$\begin{aligned} E\left(\bar{Y}_t - \bar{Y}_c\right) &= E(Y_t \mid S_i = 1) - E(Y_c \mid S_i = 0) \\ \bar{Y}_t - \bar{Y}_c &\xrightarrow{f.s.} E(Y_t \mid S_i = 1) - E(Y_c \mid S_i = 0) \end{aligned} \tag{7}$$

Damit werden von beobachteten Daten ausgehend Erwartungswerte geschätzt, die ohne weitere Annahmen nicht von „kausaler Relevanz“ (vgl., Holland, 1986) sind, da es sich bei diesen bedingten Erwartungswerten a) um den durchschnittlichen response im treatment derer, die zu t selegiert werden (bedingt an dem Ereignis $S_i = 1$) bzw. b) um den durchschnittlichen response in der Kontrollbedingung derer, die zu c selegiert werden ($S_i = 0$) handelt. Das Fundamentalproblem der kausalen Inferenz bleibt damit bei Definition und Schätzung des ACE in Folge der Selektion weiterhin bestehen, denn die bedingten Erwartungswerte $E(Y_t \mid S_i = 0)$ sowie $E(Y_c \mid S_i = 1)$ sind anhand der beobachteten Daten nicht identifizierbar sondern liegen kontrafaktisch vor und machen damit, wie folgend zu zeigen ist, eine vollständige statistische Inferenz mittels der Statistik $\bar{Y}_t - \bar{Y}_c$ auf den ACE unmöglich.

1.6.2 Identifikation des ACE

Es ist leicht aufzuweisen, dass sich der ACE additiv aus vier, an dem Ereignis $S_i = s_i$ bedingten Erwartungswerten zusammensetzt und zwei dieser bedingten Erwartungswerte stets kontrafaktisch sind. Da diese weder identifiziert noch geschätzt werden können, liefern die in (7) aufgeführten und durch $\bar{Y}_t$ bzw. $\bar{Y}_c$ erwartungstreu und konsistent schätzbaren Erwartungswerte grundsätzlich nicht alle Informationen, die für einen induktiven Schluss auf den ACE benötigt werden - das missing data-Problem bleibt weiterhin bestehen. Es folgt zunächst in

Abhängigkeit von einer probabilistischen Selektion:

$$\begin{aligned}
E(Y_t) - E(Y_c) =& \pi \left[E\left(Y_t \mid S_i = 1\right) - E\left(Y_c \mid S_i = 1\right)\right] \\
&+ (1-\pi)\left[E\left(Y_t \mid S_i = 0\right) - E\left(Y_c \mid S_i = 0\right)\right] \\
=& \left[\pi E\left(Y_t \mid S_i = 1\right) + (1-\pi) E\left(Y_t \mid S_i = 0\right)\right] - \\
& \left[\pi E\left(Y_c \mid S_i = 1\right) + (1-\pi) E\left(Y_c \mid S_i = 0\right)\right]
\end{aligned}$$

De facto sind die bedingten Erwartungswerte $E\left(Y_t \mid S_i = 0\right)$ sowie $E\left(Y_c \mid S_i = 1\right)$ anhand der Beobachtungen nicht identifizierbar, da diese a) den durchschnittlichen response im treatment derjenigen, die in die Kontrollbedingung selegiert werden bzw. b) den durchschnittlichen response in der Kontrollbedingung derjenigen, die zu t selegiert werden, messen und damit der faktischen Selektion gegenüberstehen. Um dieses *Identifikationsproblem* lösen zu können, müssen die Annahmen

$$E(Y_t \mid S_i = 1) = E(Y_t \mid S_i = 0) \text{ sowie } E(Y_c \mid S_i = 0) = E(Y_c \mid S_i = 1)$$

getroffen werden, inhaltlich demnach eine *Homogenität* der beiden Teilgruppen bezüglich ihrer potentiellen response-Variablen, so dass eine Gruppe die fehlende Information über den kontrafaktischen Erwartungswert der anderen Gruppe liefert. Unterhalb dieser Annahme folgt:

$$\begin{aligned}
E(Y_t) - E(Y_c) =& \left[\pi E\left(Y_t \mid S_i = 1\right) + (1-\pi) E\left(Y_t \mid S_i = 0\right)\right] - \\
& \left[\pi E\left(Y_c \mid S_i = 0\right) + (1-\pi) E\left(Y_c \mid S_i = 1\right)\right] \\
=& \left[\pi E\left(Y_t \mid S_i = 1\right) + (1-\pi) E\left(Y_t \mid S_i = 1\right)\right] \\
& - \left[\pi E\left(Y_c \mid S_i = 0\right) + (1-\pi) E\left(Y_c \mid S_i = 0\right)\right] \\
=& E\left(Y_t \mid S_i = 1\right) - E\left(Y_c \mid S_i = 0\right)
\end{aligned}$$

Folglich kann der ACE mit den identifizierbaren, bedingten Erwartungswerten gleichgesetzt werden und entsprechend durch die Stichprobenmittelwerte geschätzt werden; Höfler (2005, S.4) expliziert diese Homogenitätsannahme:

> „The distribution of the unobserved outcome Y_t under actual treatment c is the same as that of the observed outcome Y_t under actual treatment t; that is, under counterfactual treatment with t, the individuals actually treated with c would behave like those actually treated with t; individuals having received treatment t are substitutes for individuals having received treatment c with respect to Y_t.“

Die benötigte Homogenitätsannahme lässt sich unmittelbar mit einem unkonfundierten treatment-assignment in Verbindung bringen: Gegeben $(Y_t, Y_c) \perp\!\!\!\perp S_i$ lässt sich der Unabhängigkeit von potentiellen Variablen und Selektionsindikator folgend eine identische Verteilung der potentiellen Variablen unterhalb der Selektionsindikationen voraussetzen und die kontrafaktischen bedingten Erwartungswerte lassen sich durch die identifizierbaren ersetzen. Der Annahme $(Y_t, Y_c) \perp\!\!\!\perp S_i$ folgt zusätzlich:

$$E(Y_t \mid S_i = 1) = E(Y_t \mid S_i = 0) = E(Y_t) \text{ sowie } E(Y_c \mid S_i = 0) = E(Y_c \mid S_i = 1) = E(Y_c)$$

Die Folgen dieser Annahme sind weitreichend, denn anhand der Differenz der bedingten Erwartungswerte $E(Y_t \mid S_i = 1) - E(Y_c \mid S_i = 0)$, die durch die Stichprobenmittel $\bar{Y}_t$ und $\bar{Y}_c$ schätzbar sind, folgt eine Identifikation des ACE, so dass:

$$E(Y_t \mid S_i = 1) - E(Y_c \mid S_i = 0) \overset{(Y_t,Y_c)\perp\!\!\!\perp S_i}{=} E(Y_t) - E(Y_c)$$

Für die Schätzstatistik $\bar{Y}_t - \bar{Y}_c$ folgt einem unkonfundierten treatment-assignment und i.i.d.-Ziehungen:

$$\begin{aligned}
E\left(\bar{Y}_t - \bar{Y}_c\right) &= E\left(\frac{1}{n_t} \sum_{i:i\in\{I_i=1,S_i=1\}} y_i - \frac{1}{n_c} \sum_{i:i\in\{I_i=1,S_i=0\}} y_i\right) \\
&\overset{\text{SUTVA}}{=} E\left(\frac{1}{n_t} \sum_{i:i\in\{I_i=1,S_i=1\}} y_{i,t} - \frac{1}{n_c} \sum_{i:i\in\{I_i=1,S_i=0\}} y_{i,c}\right) \\
&= E\left(\frac{1}{n_t} \sum_{i:i\in\{I_i=1,S_i=1\}} y_{i,t}\right) - E\left(\frac{1}{n_c} \sum_{i:i\in\{I_i=1,S_i=0\}} y_{i,c}\right) \\
&= \frac{1}{n_t} \cdot n_t \cdot E\left(Y_t \mid S_i = 1\right) - \frac{1}{n_c} \cdot n_c \cdot E\left(Y_c \mid S_i = 0\right) \\
&\overset{(Y_t,Y_c)\perp\!\!\!\perp S_i}{=} E\left(Y_t\right) - E\left(Y_c\right)
\end{aligned}$$

sowie:

$$\bar{Y}_t - \bar{Y}_c \xrightarrow{f.s.} E(Y_t \mid S_i = 1) - E(Y_c \mid S_i = 0) \overset{(Y_t,Y_c)\perp\!\!\!\perp S_i}{=} E(Y_t) - E(Y_c).$$

1.6.3 Eigenschaften des Schätzers bei Randomisierung

Wie gezeigt wurde, ist eine Unabhängigkeit der potentiellen response-Variablen von der Zuweisung der Einheiten zu den Experimentalbedinungen, dem sog. *ignorable treatment-assignment*

(z.B., Holland, 1986; Morgan & Winship, 2007; Rosenbaum, 1984b) bzw. *unconfounded treatment-assignment* (z.B., Imbens & Rubin, 2012; Lunceford & Davidian, 2004), formal $(Y_t, Y_c) \perp\!\!\!\perp S_i$, vorauszusetzen, wenn der ACE als *kausaler Parameter* identifiziert und erwartungstreu sowie konsistent durch $\bar{Y}_t - \bar{Y}_c$ geschätzt werden soll.

Die Annahme, $(Y_t, Y_c) \perp\!\!\!\perp S_i$, kann bedingungslos bei Randomisierung als gültig angenommen werden: In Folge der zufälligen Generierung der Werte $S_i = s_i$ spielt die Konditionierung an S_i keine Rolle mehr; der Selektionsindikator ist unabhängig von jeglichen Werten - gemessen oder ungemessen - der Einheiten. Lunceford und Davidian (2004, S.2939) führen aus:

> „In a randomized trial, as Z is determined for each participant at random, it is unrelated how s/he might *potentially respond*, and thus $(Y_0, Y_1) \perp\!\!\!\perp Z(\ldots)$“

In Folge eines randomisierten treatment-assignments weiß i vor Realisierung $S_i = s_i$ nicht, zu welchem $k \in K$ sie selegiert wird und kann sich auch nicht aufgrund einer Kenntnis der Wirkung des treatments, d.h. auf Grund einer Kenntnis über ihren potentiellen Wert $y_{i,t}$, selber t zuordnen. Heckman (1991, S.3) führt aus:

> „Randomization ensures that there is no selection bias among participants <u>i.e.</u>, there is no selection into or out of the program on the basis of outcomes for the randomized sample.“

Es folgt bei Randomisierung stets: $\Pr(S_i = 1 \mid y_{i,c}, y_{i,t}) = \Pr(S_i = 1)$, so dass $E(Y_t \mid S_i = 1) = E(Y_t \mid S_i = 0)$ bzw. $E(Y_c \mid S_i = 0) = E(Y_c \mid S_i = 1)$ dem folgend vorausgesetzt werden kann. Damit kann durch Randomisierung stets sicher gestellt werden, dass es keine Systematik bei der Zuweisung der Einheiten gibt, so dass die resultierenden Experimentalgruppen Informationen, die durch die Selektion verloren gehen, für die andere Gruppe liefern. Des weiteren folgt aus der Unabhängigkeitsannahme unmittelbar: $E(Y_t \mid S_i = 1) = E(Y_t)$ und $E(Y_c \mid S_i = 0) = E(Y_c)$, so dass eine Identifikation des ACE sowie eine erwartungstreue und konsistente Schätzung des ACE durch die Stichprobenmittelwerte ermöglicht wird.

Zusätzlich gilt, dass bei Randomisierung auch die erhobenen Kovariablen $\mathbf{X}$ sowie die nicht erhobenen Kovariablen $\mathbf{Z}$ unabhängig von S_i sind (z.B., Holland & Rubin, 1988; Imai et al., 2008; Rosenbaum, 1995; Stuart, 2010); es gilt bei Randomisierung: $(\mathbf{X}, \mathbf{Z}) \perp\!\!\!\perp S_i$. In Konsequenz gilt für die Schätzfunktion zur Schätzung des ACE:

$$E\left(\bar{Y}_t - \bar{Y}_c \mid \vec{x}_i\right) = E(Y_t \mid S_i = 1, \vec{x}_i) - E(Y_c \mid S_i = 0, \vec{x}_i) = E(Y_t \mid \vec{x}_i) - E(Y_c \mid \vec{x}_i) = E(Y_t - Y_c \mid \vec{x}_i),$$

sowie:

$$E\left(\bar{Y}_t - \bar{Y}_c \mid \vec{z}_i\right) = E(Y_t \mid S_i = 1, \vec{z}_i) - E(Y_c \mid S_i = 0, \vec{z}_i) = E(Y_t \mid \vec{z}_i) - E(Y_c \mid \vec{z}_i) = E(Y_t - Y_c \mid \vec{z}_i).$$

Der Unabhängigkeit vom Selektionsindiktor folgt eine identische Verteilung der Kovariablen, so dass $F(\mathbf{X}, \mathbf{Z} \mid S_i = 1) = F(\mathbf{X}, \mathbf{Z} \mid S_i = 0)$. In Konsequenz heißt dies, dass sämtlich relevante Variablen, die zu einer Konfundierung führen könnten, stochastisch unabhängig von S_i sind und die resultierenden Experimentalgruppen direkt in ihren potentiellen response-Werten sowie den Kovariablen vergleichbar sind.

Insbesondere in der Ökonometrie finden sich Kritiker, die die Randomisierung als Selektionsmethode ablehnen, da diese eine Verzerrung bei der Schätzung des treatment-Effekts evozieren kann, die losgelöst ist von jeglichen vorherigen Überlegungen. In einer Vielzahl von Veröffentlichungen wird von einem sog. *randomization bias* ausgegangen (z.B., Heckman, 1991; Heckman & Smith, 1995; Imai et al., 2008), der sich in einem ein veränderten Verhalten derer, die prinzipiell gewillt sind, an einer Studie teilzunehmen, durch den Zwischenschritt der zufälligen Selektion äußert. Heckman (1991, S.17) führt aus:

> „If individuals who might have enrolled in a nonrandomized regime make plans anticipating enrollment in training, adding uncertainty at the acceptance stage may alter their decision to apply or to undertake activities complementary to training. Risk averse persons will be eliminated from the program.“

1.6.4 Eigenschaften des Schätzers in Observational Studies

Die Schätzung des ACE in observational studies beginnt stets mit der Beobachtung der Werte $(y_i, s_i, \vec{x}_i)$ für i bei Datenerhebung; der entscheidende Unterschied zum vorherigen Falle ist die bereits vorliegende Selektion von i in eine der beiden Experimentalbedingungen. Da die Selektion nicht kontrolliert und randomisiert stattfindet, muss grundsätzlich angenommen werden, dass das treatment-assignment konfundiert sein kann und $(Y_t, Y_c) \perp\!\!\!\perp S_i$ verletzt ist. Folglich lassen sich die vorherig getroffenen Annahmen, $E(Y_t \mid S_i = 1) = E(Y_t \mid S_i = 0) = E(Y_t)$ sowie $E(Y_c \mid S_i = 0) = E(Y_c \mid S_i = 1) = E(Y_c)$, deren Gültigkeit auf dem unkonfundierten treatment-assignment beruhten und eine Identifikation des ACE ermöglichten, nicht mehr aufrecht erhalten; stattdessen können mit der Selektion Unterschiede in den Verteilungen der potentiellen Werte zwischen den Gruppen einhergehen; es liegt damit ein sog. *selection bias*

vor. Die möglichen Konsequenzen sollen folgend durch eine algebraische Zerlegung des ACE aufgewiesen werden: es lässt sich unmittelbar zeigen, dass der ACE nicht mehr mit der Differenz $E(Y_t \mid S_i = 1) - E(Y_c \mid S_i = 0)$ gleichgesetzt werden kann. Stattdessen folgen dem konfundierten treatment-assignment zusätzliche Effekte, die sich additiv auf den ACE legen, so dass in Folge die Statistik $\bar{Y}_t - \bar{Y}_c$ verzerrt und inkonsistent den ACE schätzt:

$$
\begin{aligned}
\underbrace{E(Y_t) - E(Y_c)}_{e} = & \quad [\pi \underbrace{E\left(Y_t \mid S_i = 1\right)}_{a} + (1-\pi) \underbrace{E\left(Y_t \mid S_i = 0\right)}_{b}] - \\
& \quad [\pi \underbrace{E\left(Y_c \mid S_i = 1\right)}_{c} + (1-\pi) \underbrace{E\left(Y_c \mid S_i = 0\right)}_{d}] \\
e = & \quad [\pi \cdot a + (1-\pi) \cdot b] - [\pi \cdot c + (1-\pi) \cdot d] \\
e = & \quad [\pi \cdot a + (1-\pi) \cdot b] - [\pi \cdot c + (1-\pi) \cdot d] \pm a \\
e = & \quad \pi \cdot a + b - \pi \cdot b - \pi \cdot c - d + \pi \cdot d + a - a \\
a - d = & \quad e - \pi \cdot a - b + \pi \cdot b + \pi \cdot c - \pi \cdot d + a \\
a - d = & \quad e - \pi \cdot a - b + \pi \cdot b + \pi \cdot c - \pi \cdot d + a \pm (c - d) \\
a - d = & \quad e + a \cdot (1-\pi) - b \cdot (1-\pi) + \pi \cdot c - \pi \cdot d \pm (c - d) \\
a - d = & \quad e + (c - d) + a \cdot (1-\pi) - b \cdot (1-\pi) + \pi \cdot c - \pi \cdot d - c + d \\
a - d = & \quad e + (c - d) + a \cdot (1-\pi) - b \cdot (1-\pi) - c \cdot (1-\pi) + d \cdot (1-\pi) \\
a - d = & \quad e + (c - d) + (1-\pi) \cdot (a - b - c + d) \\
a - d = & \quad e + (c - d) + (1-\pi) \cdot (a - c - b + d) \\
a - d = & \quad e + (c - d) + (1-\pi) \cdot ((a - c) - (b - d))
\end{aligned}
$$

Durch Substitution folgt:

$$
\begin{aligned}
E(Y_t \mid S_i = 1) - E(Y_c \mid S_i = 0) = & \quad (E(Y_t) - E(Y_c)) + (E(Y_c \mid S_i = 1) - E(Y_c \mid S_i = 0)) \\
& + (1-\pi) \cdot ((E(Y_t \mid S_i = 1) - E(Y_c \mid S_i = 1)) \\
& - (E(Y_t \mid S_i = 0) - E(Y_c \mid S_i = 0)))
\end{aligned}
$$

Festzuhalten ist eine fast sichere Konvergenz der Statistik $\bar{Y}_t - \bar{Y}_c$ gegen $E(Y_t \mid S_i = 1) - E(Y_c \mid S_i = 0)$; entgegen einem randomisierten treatment-assignment kann jedoch diese Differenz nicht gleichgesetzt werden mit dem ACE. Stattdessen treten mehrere Möglichkeiten einer Verzerrung in der Gleichung auf, die folgend erläutert werden sollen:

- $E(Y_c \mid S_i = 1) - E(Y_c \mid S_i = 0)$ beschreibt einen *baseline bias*, der sich in einer gemessenen Differenz zwischen beiden Gruppen äußert, die lediglich durch einen Unterschied im Grundniveau, d.h. in Abwesenheit des treatments, beider Gruppen zustande kommt.

- $(E(Y_t \mid S_i = 1) - E(Y_c \mid S_i = 1)) - (E(Y_t \mid S_i = 0) - E(Y_c \mid S_i = 0))$ entspricht einem *differentiellen treatment-Effekt*, der sich dadurch äußert, dass Einheiten, die sich dem treatment zuordnen, auf dieses abweichend reagieren als Einheiten, die sich der Kontrollbedingung selegieren.

1.6.5 Konditionierte Treatment-Effekte

Es lassen sich ausgehend von der Verletzung der Annahme $(Y_t, Y_c) \perp\!\!\!\perp S_i$ zwei weitere treatment-Effekte definieren, gegen die die Statistik $\bar{Y}_t - \bar{Y}_c$ unter Gültigkeit zusätzlicher Annahmen konvergiert, da diese, wie aufgewiesen, nicht gegen den ACE konvergiert:

Insofern $E(Y_c \mid S_i = 1) \neq E(Y_c \mid S_i = 0)$, die Annahme $E(Y_t \mid S_i = 1) = E(Y_t \mid S_i = 0)$ jedoch weiterhin aufrecht erhalten werden kann, konvergiert $\bar{Y}_t - \bar{Y}_c$ gegen den treatment-Effekt derer, die nicht dem treatment zugewiesen worden sind:

$$\begin{aligned} \bar{Y}_t - \bar{Y}_c \xrightarrow{f.s.} \quad & E(Y_t \mid S_i = 1) - E(Y_c \mid S_i = 0) \quad \mid \text{Da: } E(Y_t \mid S_i = 1) = E(Y_t \mid S_i = 0) \\ = \quad & E(Y_t \mid S_i = 0) - E(Y_c \mid S_i = 0) \end{aligned}$$

Es handelt sich hierbei um den *average treatment effect of the untreated* mit

$$E(Y_t \mid S_i = 0) - E(Y_c \mid S_i = 0) = \frac{1}{N_c} \sum_{i:S_i=0} \tau_i \tag{8}$$

Insofern $E(Y_t \mid S_i = 1) \neq E(Y_t \mid S_i = 0)$, jedoch die Annahme $E(Y_c \mid S_i = 0) = E(Y_c \mid S_i = 1)$ weiterhin aufrecht erhalten werden kann, so ist der Schätzer $\bar{Y}_t - \bar{Y}_c$ konsistent für den *average treatment effect of the treated*. Es gilt:

$$\begin{aligned} \bar{Y}_t - \bar{Y}_c \xrightarrow{f.s.} \quad & E(Y_t \mid S_i = 1) - E(Y_c \mid S_i = 0) \quad \mid \text{Da: } E(Y_c \mid S_i = 0) = E(Y_c \mid S_i = 1) \\ = \quad & E(Y_t \mid S_i = 1) - E(Y_c \mid S_i = 1) \end{aligned}$$

mit

$$E(Y_t \mid S_i = 1) - E(Y_c \mid S_i = 1) = \frac{1}{N_t} \sum_{i:S_i=1} \tau_i \tag{9}$$

Die aufgeführten bedingten treatment-Effekte sind insbesondere in der Ökonometrie und Biometrie von Interesse und stellen *de facto* theoretische Größen dar, die zumeist aus dem inhaltlichen Kontext heraus definiert werden (z.B., Heckman, 1991; Imbens & Angrist, 1994; Morgan, 2001). Ein Beispiel sei hier die Evaluierung der Frage, wie sich eine Asbestexposition auf die Gesundheit auswirkt - der Fokus der Untersuchung richtet sich hier lediglich auf diejenigen Einheiten, die entsprechend expositioniert worden sind ($S_i = 1$) und nicht auf eine allgemeine Grundgesamtheit, die entsprechend Einheiten umfasst, die nicht Asbest ausgesetzt waren. Wie an späterer Stelle gezeigt wird, liegen Schätzstatistiken, wie beispielsweise das *matching*, vor, die lediglich eine Definition des ATT zulassen. Der Fokus der vorliegenden Arbeit beruht primär auf dem in (4) definierten ACE, Abweichungen werden entsprechend kenntlich gemacht.

1.7 Kausale Inferenz in Experimenten

1.7.1 Taxonomie von randomisierten Experimenten

Wie bereits aufgeführt, ist ein randomisiertes treatment-assignment grundlegend dadurch charakterisiert, dass a-priori die Selektionswahrscheinlichkeit $\Pr(S_i = 1)$ auf einen festen Wert π, $\pi \in (0, 1)$, festgelegt wird. In Konsequenz folgt, dass

a) die Wahrscheinlichkeit, zu t selegiert zu werden, bekannt und für alle Einheiten identisch ist,

b) die Wahrscheinlichkeit für alle Einheiten, ins treatment zu gelangen, positiv ist,

c) alle realisierbaren Vektoren $\vec{S} = \vec{s}$, die jeweilig eine mögliche Selektion der N Einheiten zusammenfassen, in Form des Ergebnisraum $\Omega_{\vec{S}}$ bekannt sind.

Explizit der letzte Punkt soll in diesem Kapitel verdeutlicht werden, da er die Grundlage für den *Fisher'schen* Lösungsansatz des Fundamentalproblems darstellt: Einhergehend mit einer vollzogenen Selektion der N Einheiten in beide Bedingungen wird der Vektor $\vec{s}$ realisiert, dessen N Einträge aus den Realisierungen $S_i = s_i$ bestehen, d.h.:

$$\vec{s} := \begin{pmatrix} s_1 \\ s_2 \\ \vdots \\ s_N \end{pmatrix} \quad i = 1, \ldots, N.$$

In Folge der Randomisierung lassen sich zu dem realisierten $\vec{s}$ der Ergebnisraum $\Omega_{\vec{S}}$, der alle möglichen $\vec{S} = \vec{s}$ als Elementarereignisse zusammenfasst, sowie die zugehörigen Wahrscheinlichkeiten $\Pr(\vec{S} = \vec{s})$ für ein konkretes treatment-assignment angeben. Dies gilt es hervorzuheben: Randomisierung ermöglicht stets, sämtliche möglichen Selektionsmuster, $\vec{s} \in \Omega_{\vec{S}}$, die jedoch nicht realisiert worden sind, anzugeben und sämtliche zufällige Variation, die sich in Folge der Randomisierung einstellt, durch Ermittlung der Wahrscheinlichkeit $\Pr(\vec{S} = \vec{s})$ kalkulierbar zu machen. Folglich handelt es sich bei $\Omega_{\vec{S}}$ als Menge aller Elementarereignisse um die Zusammenfassung aller der Randomisierung folgenden Zuweisungsmöglichkeiten der N Einheiten auf die Experimentalbedingungen.

Wie folgend zu zeigen ist, lassen sich grundsätzliche Unterschiede bei der Zuweisung einer positiven Wahrscheinlichkeit für ein konkretes $\vec{s}$ treffen, so dass eine *Taxonomie von randomisierten Experimenten* eingeführt werden soll:

Im *einfachsten Falle* lässt sich die Selektion der N Einheiten als ein $N-$fach durchgeführter *Bernoulli-Versuch* verstehen, so dass $S_i \sim B(1, \pi)$ als Modell für die Selektion von i und entsprechend

$$\Pr\left(\vec{S} = \vec{s}\right) = \prod_{i=1}^{N} \pi^{s_i}(1-\pi)^{1-s_i} = \pi^{N_t}(1-\pi)^{N_c}, \quad \pi \in (0,1) \tag{10}$$

gilt. Daraus folgt, dass $\Omega_{\vec{S}} = \{0,1\}^N$ und $\mid \Omega_{\vec{S}} \mid = 2^N$. Entscheidend damit: Die Anzahl derer, die zu t selegiert werden soll, ist nicht festgelegt und es können $\vec{s}$ resultieren, die eine vollständige Selektion aller N Einheiten in eine der beiden Bedingung angeben.

In praxi wird dieses Problem dadurch umgangen, dass die Anzahl derer, die zu t selegiert werden sollen, mit $0 < N_t < N$ festgelegt wird und entsprechend $\Pr(S_i = 1) = \pi = \frac{N_t}{N}$ für alle N Einheiten gilt. Dieses Vorgehen findet sich als *completely randomized experiment* (z.B., Imbens & Rubin, 2012; Rosenbaum, 1995) und führt zu der Restriktion:

$$\Omega_{\vec{S}} := \left\{ \vec{s} \,\middle|\, \sum_{i=1}^{N} S_i = N_t \right\}, \mid \Omega_{\vec{S}} \mid = \binom{N}{N_t}.$$

Damit folgt für die Wahrscheinlichkeit eines einzelnen treatment-assignments:

$$\Pr\left(\vec{S} = \vec{s}\right) = \begin{cases} \frac{1}{\binom{N}{N_t}}, & \text{wenn } \sum_{i=1}^{N} S_i = N_t \\ 0 & \text{sonst.} \end{cases} \tag{11}$$

Diese Restriktion führt damit stets zu realisierten treatment-assignments, innerhalb derer die selbe Anzahl an Einheiten zu t selegiert wird und unplausible $\vec{s}$ nicht realisiert werden können.

Das Konzept des *completely randomized experiment* lässt sich ausweiten auf *stratifizierte randomisierte Experimente* (z.B., Fisher, 1926; Imbens & Rubin, 2012; Rosenbaum, 1995), die im einfachsten Fall a-priori versuchen, den möglichen Störeffekt einer Variablen X mit dem treatment-assignment einhergehend zu balancieren. Hierzu werden alle N Einheiten zunächst *strata* zugeordnet, die, im diskreten Fall, mit der Ausprägung $X = x_j$, $j = 1, \ldots, v$, einhergehend definiert sind. In jedem stratum j lassen sich damit N_j Einheiten beobachten, die der Stratifikation folgend randomisiert den Bedingungen zugeteilt werden.

Mit $0 < N_{tj} < N_j$ folgt damit: $\Pr(S_{ij} = 1) = \frac{N_{tj}}{N_j}$ und entsprechend

$$\Pr\left(\vec{S} = \vec{s}\right) = \begin{cases} \prod_{j=1}^{v} \frac{1}{\binom{N_j}{N_{tj}}}, & \text{wenn } \sum_{i:X=x_j}^{N_j} S_{ij} = N_{tj}, j = 1, \ldots, v \\ 0 & \text{sonst.} \end{cases} \tag{12}$$

Den drei Klassifikationsmöglichkeiten gemeinsam ist die bekannte Wahrscheinlichkeit für die Beobachtung eines treatment-assignments $\vec{s}$. Explizit Fisher (1926) nutzte diese bekannten Wahrscheinlichkeiten für den Aufbau eines Hypothesentests, um die *sharp null hypothesis*, die vollzogene Intervention sei ohne Effekt, nach Durchführung des Experimentes zu testen. Wie zu zeigen ist, stellt das Fisher'sche Vorgehen eine mögliche Inferenz auf den individuellen treatment-Effekt τ_i dar, die dem Neyman'schen Vorgehen, welches den ACE fokussierte, gegenüber steht und die ausschließlich auf dem bekannten Ergebnisraum $\Omega_{\vec{S}}$ beruht. Wie zu zeigen ist, lässt sich die Verteilung einer möglichen Teststatistik, $T(\vec{y}, \vec{S})$, unterhalb der H_0 angeben, die vollständig auf $\Omega_{\vec{S}}$ beruht und deren Wahrscheinlichkeitsverteilung unterhalb der H_0 durch $\Omega_{\vec{S}}$ determiniert ist (z.B., Imbens & Rubin, 2012; Rosenbaum, 1995).

1.7.2 Randomisierte Experimente - Vorgehen nach Fisher

Fisher (1926) interessierte sich wie Neyman (1932) für den Nachweis des dem treatment immanenten Effektes und sah die Lösung in dem Aufbau eines Hypothesentests, der einerseits auf den beobachteten y_i-Werten, andererseits auf den bekannten Wahrscheinlichkeiten $\Pr(\vec{S} = \vec{s})$ beruht. Eine zu testende H_0 nach Durchführung eines Experimentes geht an dieser Stelle einher mit der Annahme, dass die Durchführung von t im Verhältnis zu c keinen Unterschied in den potentiellen response-Werten, $y_{i,t}$ und $y_{i,c}$, induziert, sondern unter jeder Experimentalbedingung die potentiellen response-Werte von i konstant bleiben; erwartet wird unter der H_0:

$$H_0: \; y_{i,t} = y_{i,c} \quad i = 1, \ldots, N.$$

Diese *sharp null hypothesis* formuliert damit eine statistische Hypothese, die sich direkt auf die potentiellen response-Werte der Einheit i bezieht und damit keine Annahmen über einen globalen Parameter wie den ACE formuliert. Der Nullhypothese nach gilt damit für den beobachteten response von i: $y_i = y_{i,t} = y_{i,c}$.

Dem Experiment folgend lassen sich alle N beobachteten y_i-Werte als Vektor $\vec{y}$ zusammenfassen und unter Gültigkeit der SUTVA lässt sich der y_i-Wert gleichsetzen mit demjenigen potentiellen response-Wert von i, der der Selektion in eines der $k \in K$ folgt, so dass:

$$\vec{y} := \begin{pmatrix} y_{1,k} \\ y_{2,k} \\ \vdots \\ y_{N,k} \end{pmatrix}.$$

Der H_0 folgend ist es irrelevant, in welche Bedingung i selegiert wurde, da unterhalb der anderen Bedingung derselbe y_i-Wert beobachtet werden würde. Damit trifft die H_0 konkrete Annahmen über den *missed* potentiellen response-Wert nach Selektion und stellt ein einfaches *Imputationsverfahren* für die fehlenden Werte dar, wie auch Imbens und Rubin (2012, S.60) betonen:

> „(...) under the sharp null hypothesis, all the missing values can be inferred from the observed ones.“

Um die H_0 testen zu können, lässt sich der bekannte Ergebnisraum $\Omega_{\vec{S}}$ nutzen, um die Wahrscheinlichkeitsverteilung einer beliebigen Prüfgröße $T(\vec{y}, \vec{S})$ aufzustellen, wie z.B. der *treatment-Summe*, $T(\vec{y}, \vec{S}) := \vec{y}\,{}^t\vec{S}$. Es ist unterhalb der H_0 eindeutig, dass die Wahrscheinlichkeitsverteilung der Prüfgröße vollständig durch die Wahrscheinlichkeiten $\Pr(\vec{S} = \vec{s})$ bestimmt ist, da unterhalb aller möglichen $\vec{S} = \vec{s}$ und unter Gültigkeit der H_0 $\vec{y}$ identisch formuliert wird. Somit lassen sich die Realisierungen der Prüfgröße, $t(\vec{y}, \vec{S})$, ermitteln, indem jedes $\vec{s} \in \Omega_{\vec{S}}$ in die Prüfgröße $T(\vec{y}, \vec{S})$ eingesetzt wird, so dass der als konstant angenommene Vektor $\vec{y}$ zur Ermittelung der alternativ möglichen Werte der Prüfgröße genutzt wird. Auf Grundlage der resultierenden Wahrscheinlichkeitsverteilung der Prüfgröße kann folglich der $p-$Wert, d.h., die Wahrscheinlichkeit, den beobachteten Wert der Prüfgröße oder extremere unterhalb der H_0 zu finden, bestimmt werden und damit abglichen werden, ob:

$$\Pr\Big(T(\vec{y}, \vec{S})) \geq t(\vec{y}, \vec{S}))\Big) \leq \alpha,$$

wobei α : a-priori festgelegte Wahrscheinlichkeit, die H_0 abzulehnen, wenn diese als gültig angenommen wird (sog. *Signifikanzniveau*).

Ein Beispiel soll die Logik dieses Hypothesentests verdeutlichen: Angenommen, in einem randomisierten Experiment mit $N = 4$ und $N_t = 2$ wurden folgendes treatment-assignment $\vec{s}$ und folgender response-Vektor $\vec{y}$ beobachtet:

$$\vec{y} := \begin{pmatrix} 4 \\ 0 \\ 2 \\ 3 \end{pmatrix}, \quad \vec{s} := \begin{pmatrix} 1 \\ 0 \\ 0 \\ 1 \end{pmatrix}.$$

Den vorherigen Überlegungen folgt: $\Pr(\vec{S} = \vec{s}) = \frac{1}{\binom{4}{2}} = \frac{1}{6}$. Ausgehend von der als gültig angenommenen Nullhypothese, $H_0 : y_{i,t} = y_{i,c}, i = 1, \ldots, 4$, lässt sich diesen Annahmen folgend die Verteilung einer beliebigen Teststatistik konstruieren; folgend wird die *treatment-Summe* als Teststatistik gewählt:

$$t(\vec{y}, \vec{S}) = \vec{s}^{\,t} \cdot \vec{y} = (1\ 0\ 0\ 1) \begin{pmatrix} 4 \\ 0 \\ 2 \\ 3 \end{pmatrix} = 7$$

Unter der H_0 wären weitere treatment-Summen möglich gewesen, die sich lediglich dadurch ermitteln lassen, als dass die möglichen $\vec{s} \in \Omega_{\vec{S}}$ mit dem bestehenden response-Vektor $\vec{y}$ zur Berechnung der treatment-Summe genutzt wird; es ergibt sich damit als Träger der Teststatistik: $\mathbb{T} := \{2, 3, 4, 5, 6, 7\}$, wobei für jede der Realisierungen gilt: $\Pr(T(\vec{y}, \vec{S}) = t(\vec{y}, \vec{S})) = \frac{1}{6}$. Demnach liegt unter der H_0 die Wahrscheinlichkeit, den Wert der Teststatistik zu beobachten, bei $\Pr(T(\vec{y}, \vec{S}) = 7) = \frac{1}{6}$. Mit einem festgelegten Signifikanzniveau α, z.B. $\alpha = 0.2$, welches den Ablehnbereich unterhalb der H_0 definiert, könnte an dieser Stelle die Nullhypothese abgelehnt werden, denn es gilt in diesem Falle:

$$\Pr(T(\vec{y}, \vec{S})) \geq t(\vec{y}, \vec{S}))) \leq \alpha = \frac{1}{6} \leq 0.2$$

1.8 Observational Studies

Analog zu randomisierten Experimenten zielen *observational studies* auf die Messung des Effektes eines applizierten treatments ab; damit unterscheiden sich observational studies in ihrem Erkenntnisinteresse nicht von randomisierten Experimenten, jedoch lässt sich eine grundlegende Differenz zwischen beiden Designs feststellen: in observational studies werden die statistischen Einheiten nicht randomisiert den verschiedenen Experimentalbedingungen zugeführt, sondern diese sind mit der Datenerhebung den Versuchsbedingungen zugeordnet (Rosenbaum, 1995). Cochran und Chambers (1965, S.234 ff.) führen die grundlegende Idee aus:

> „The objective is to elucidate cause-and-effect relationships (...). In controlled experimentation the investigator decides on the procedures or treatments whose effects he wishes to compare, and takes steps to apply them (...). In observational studies the investigator, having decided on the types of comparison that he would like to make, often has to search for some environment in which it may be possible to collect data that provide such comparisons (...). In controlled experiments, the skillful use of randomization protects against most types of bias arising from disturbing variables (...). In observational studies, in which no random assignment of subjects to comparison groups is possible, blocking and adjustment take on the additional role of protecting against bias.“

In den Humanwissenschaften ist die Durchführung von observational studies nicht unüblich und es lassen sich in jeglichen der zugehörigen Disziplinen Beispiele finden, in denen die Ergebnisse von observational studies kausal interpretiert werden[4]. Wie zu zeigen ist, beruht die Popularität von observational studies auf der Möglichkeit, den ACE als kausalen Parameter auch dann identifizieren zu können, wenn ein vollständig randomisiertes treatment-assigment nicht möglich ist; es müssen hierzu jedoch grundlegende Modellannahmen, die der Randomisierung nahestehen, getroffen werden.

1.8.1 Definition: Observational Studies

Cochran und Chambers (1965) führen in ihrem Artikel auf, welche grundlegenden Probleme beim Umgang mit observational studies - an dieser Stelle jedoch noch in weitgreifender Hinsicht

[4]Eine Übersicht über eine Vielzahl von insbesondere sozialwissenschaftlichen Beispielen bietet das Kapitel 1.3 in Morgan und Winship (2007) sowie das Kapitel 1.2 in Rosenbaum (1995).

definiert - auftreten können, wenn diese genutzt werden, um kausale Hypothesen zu überprüfen. Es lassen sich basierend auf diesem Artikel mehrere Gründen aufführen, derer wegen observational studies im Verhältnis zu randomisierten Experimenten einen Verlust an *interner Validität* (Campbell, 1957) erleiden: so werden Studien als observational studies kategorisiert, deren Designs eine natürlich vorkommende Variation der unabhängigen Variablen und dadurch sich ergebende Assoziationen mit der abhängigen Variablen zur kausalen Interpretation der Befunde nutzen. Im Gegensatz dazu definiert in randomisierten Experimenten der Forscher das treatment als eine Handlung, dessen Effekt auf die abhängige Variable erfasst werden soll, und ist in der Lage - beispielsweise durch unterschiedliche Dosierungen oder in Relation zu einer Kontrollgruppe - eine Kontrolle über die Variation in der unabhängigen Variablen zu erlangen; in Folge dessen werden in Abhängigkeit zur kontrollierten Variation differentielle Effekte gemessen und entsprechend wird die manipulierte Variable als *exogen* angesehen (Heckman, 1991). Folgend werden in Anlehnung an Rosenbaum (1995, 2010) in dieser Arbeit nur Studiendesigns als observational studies eingegrenzt, die vergleichbar zu randomisierten Experimenten Aussagen über ein treatment treffen wollen - demnach Studien, die sich *grundlegend immer* in ein randomisiertes Experiment überführen ließen und eine Kontrolle über die unabhängige Variable gewährleisten. Rosenbaum (1995, S.1 f.) führt entsprechend aus:

> „An observational study concerns treatments, interventions, or policies and the effects they cause, and in this respect it resembles an experiment. A study without a treatment is neither an experiment nor an observational study.“

Mit dieser Eingrenzung einhergehend stehen sich randomisierte Experimente und observational studies dadurch gegenüber, als dass in observational studies der Selektionsmechanismus der Einheiten zu den Experimentalbedingungen nicht vom Forscher kontrolliert wird und die statistischen Einheiten bei Datenerhebung den Experimentalbedingungen bereits zugeordnet sind; die Datenerhebung beginnt in observational studies stets mit der Beobachtung der Werte $(y_i, s_i, \vec{x}_i)$ für die Einheit i. Folglich ist der stochastische Prozess, der die Realisierungen $S_i = s_i$ generiert, unbekannt, nicht kontrolliert, und die Selektionswahrscheinlichkeit $\Pr(S_i = 1)$ nicht notwendigerweise ein konstanter Wert. Da die statistischen Einheiten sich selbstständig den Versuchsbedingungen zuordnen können, muss damit unterstellt werden, dass das Design konfundiert ist, d.h., dass $(Y_t, Y_c) \perp\!\!\!\perp S_i$ nicht hält, und der ACE in Konsequenz basal weder identifizierbar noch konsistent durch $\bar{Y}_t - \bar{Y}_c$ schätzbar ist (vgl. Abschnitt 1.6.1). Lunceford und

Davidian (2004, S.3) verdichten vorherige Überlegungen:

> „However, in an observational study, because treatment exposure Z is not controlled, Z may not be independent of (Y_0, Y_1); indeed, the same characteristics that lead an individual to be exposed to a treatment may also be associated or "confounded," with his/her response."

Als Lösung für dieses Problem wird folgend in observational studies als Modell eine Konfundierung durch die in $\mathbf{X}$ zusammengefassten Charakteristika unterstellt, die dem Modell nach entsprechende Störvariablen darstellen können. Wie zu zeigen ist, lässt sich ausgehend von diesen Überlegungen *als Modell* in observational studies eine probabilistische Selektion, die vollständig in Abhängigkeit der Kovariablenwerte $\vec{x}_i$ gestellt wird, formulieren, so dass in Konsequenz bedingt an $\vec{x}_i$ das treatment-assignment als randomisiert angesehen werden kann. Dem folgend kann an $\vec{x}_i$ bedingt eine Unkonfundierung unterstellt werden und die Verletzung der Annahme $(Y_t, Y_c) \perp\!\!\!\perp S_i$ lässt sich bedingt an $\vec{x}_i$ lösen, das Design ist damit *strongly ignorable* (Rosenbaum, 1984b).

Somit wird als Modell folgend unterstellt, dass i bedingt an $\vec{x}_i$ durch einen exemplarischen Münzwurf zum treatment selegiert wird und sich diese probabilistische Selektion gegeben $\vec{x}_i$ durch die bedingte Wahrscheinlichkeit $\Pr(S_i = 1 \mid \vec{x}_i)$, $0 < \Pr(S_i = 1 \mid \vec{x}_i) < 1$, formalisieren lässt, die folgend als propensity score definiert ist und das treatment-assignment in observational studies bestimmt (z.B., Rosenbaum, 1984b, 1995).

Unter der Annahme, dass die Selektionswahrscheinlichkeit allein durch $\vec{x}_i$ bedingt wird, hält folglich $\mathbf{X} \perp\!\!\!\perp S_i$ nicht mehr: Unterhalb eines solchen Modells ordnen sich die Versuchspersonen auf Grundlage bestimmter persönlicher Merkmale selbstständig den Bedingungen zu. Die Konsequenzen eines von den Kovariablen abhängigen treatment-assignments sind weitreichend und unintuitiv, da die Kovariablen einen Einfluss auf die potentiellen-Werte $(y_{i,t}; y_{i,c})$ ausüben können (vgl. (3)), so dass dem Selektionsmodell folgend:

$$E(Y_i \mid S_i = s_i, \vec{x}_i) \neq E(Y_i \mid S_i = s_i)$$

unterstellt werden muss. An dieser Stelle fungieren die Kovariablen als *confunders*: Sie bedingen einerseits die Selektionswahrscheinlichkeit, so dass in Folge ein *overt bias*, mit $F(\mathbf{X} \mid S_i = 1) \neq F(\mathbf{X} \mid S_i = 0)$, resultiert - die sind Gruppen nicht mehr homogen in ihren Kovariablen -, andererseits determinieren sie die potentiellen-Werte (vgl. Abb. 1). Wie zu zeigen ist, führt je-

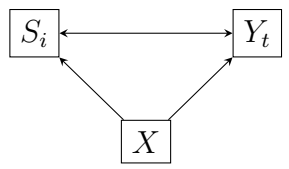

Abbildung 1: Modell einer Störvariablen X

doch die Vorstellung, dass bedingt an $\vec{x}_i$ das treatment-assignment randomisiert erfolgt und die Selektionswahrscheinlichkeit vollständig durch $\vec{x}_i$ erklärt wird, zu einem sog. *strongly ignorable treatment-assignment*: Basal halten die Annahmen $\mathbf{X} \perp\!\!\!\perp S_i$ sowie $(Y_t, Y_c) \perp\!\!\!\perp S_i$ nicht, *konditioniert* an $\vec{x}_i$ halten jedoch $(Y_t, Y_c) \perp\!\!\!\perp S_i \mid \vec{x}_i$ sowie $\mathbf{X} \perp\!\!\!\perp S_i \mid \vec{x}_i$ - das Design ist damit bedingt an $\vec{x}_i$ unkonfundiert, so dass bedingt an $\vec{x}_i$ die response-Variablen unabhängig vom Selektionsindikator sind, und die Gruppen sind bedingt an $\vec{x}_i$ in den konfundierenden Störvariablen direkt vergleichbar. Dies gilt es folgend zu formalisieren.

1.8.2 Modell: Treatment-Assignment in Observational Studies

Overt Bias / Selection On Observables

Ein von $\vec{x}_i$ abhängiges treatment-assignment, wie es als Selektionsmodell in observational studies vorgeschlagen wird, schlägt sich nieder in einem *overt bias*, einer Differenz der Verteilungen der beobachteten Kovariablen zwischen den Experimentalbedingungen, formal: $F(\mathbf{X} \mid S_i = 1) \neq F(\mathbf{X} \mid S_i = 0)$ (Rosenbaum, 1995). Zum Teil wird das folgend zu definierende Selektionsmodell in ökonometrischer und biometrischer Literatur als *selection on observable* (Heckman, 1990) klassifiziert.

Unterhalb der Modellannahme, dass gemessene Kovariablenwerte, $\vec{x}_i$, die Selektion von i zu t bedingen, muss damit unterstellt werden, dass diese Kovariablen Störvariablen darstellen, die entsprechend die potentiellen response-Werte determinieren (vgl. (3)); entsprechend ist das Design durch $\vec{x}_i$ konfundiert. Lässt sich jedoch die Selektion gegeben $\vec{x}_i$ als randomisiert auffassen, folgt, dass das Design - konditional an $\vec{x}_i$ - nicht mehr konfundiert und der ACE bedingt an $\vec{x}_i$ sowohl identifizierbar als auch schätzbar ist. Lunceford und Davidian (2004, S. 4) führen aus:

„In an observational study, although $(Y_0, Y_1) \perp\!\!\!\perp Z$ is unlikely to hold, it may be possi-

ble to identify subject characteristics related both potential response and treatment exposure, referred to as "confunders". If we believe that $\mathbf{X}$ contains all such confounders, then, for individuals sharing a particular value of $\mathbf{X}$, there would be no association between the exposure states and the value of potential responses, i.e. treatment exposure among individuals with a particular $\mathbf{X}$ is essentially at random. Formally, Y_0, Y_1 are independent of treatment exposure conditional on $\mathbf{X}$, written $(Y_0, Y_1) \perp\!\!\!\perp Z \mid \mathbf{X}$."

Damit gilt es folgend, ein probabilistisches treatment-assignment der Einheit i in Abhängigkeit des Kovariablenvektor $\vec{x}_i$ zu formulieren, so dass bedingt an $\vec{x}_i$ eine Randomisierung zu Grunde gelegt werden kann. Die Gültigkeit anschließender Überlegungen bezüglich der Identifikation und Schätzung des ACEs beruht vollständig auf einem treatment-assignment, das durch $\vec{x}_i$ erklärt wird und bedingt an $\vec{x}_i$ als randomisiert aufgefasst werden kann; ein Einfluss der nicht gemessenen Kovariablenwerte $\vec{z}_i$ sowie der potentiellen response-Werte $(y_{i,t}; y_{i,c})$ wird dem Modell nach auf die Selektionswahrscheinlichkeit ausgeschlossen (vgl. Rosenbaum, 2010, S.66). Durch das folgend formalisierte Modell eines treatment-assignments in observational studies soll damit verdeutlicht werden, dass die Wahrscheinlichkeit, dem treatment zugewiesen zu werden,

a) kein konstanter Wert für alle statistischen Einheiten ist,

b) vollständig durch $\vec{x}_i$ determiniert wird,

c) als konstant gegeben einem $\vec{x}_i$ angenommen wird und

d) als unabhängig von den unbeobachteten Kovariablen $\vec{z}_i$ sowie den potentiellen response-Werten $y_{i,t}$ und $y_{i,c}$ modelliert wird.

Rosenbaum (2010, S.65) führt exemplarisch den Kern des Modells durch die *grundlegende Vorstellung* eines *probabilistischen* und *individualisitic treatment-assignment* in observational studies auf:

„(...) we imagine that subject l received treatment with probability π_l, independently of other subjects, where π_l may vary from one person to the next and is not known."

Statistisch formuliert bedeutet dies zunächst als Grundlage für das Selektionsmodell:

$$\begin{aligned}&\Pr(S_1 = s_1, \ldots, S_N = s_N \mid y_{t,1}, y_{c,1}, \vec{x}_1, \vec{z}_1, \ldots, y_{t,N}, y_{c,N}, \vec{x}_N, \vec{z}_N)\\ &= \prod_{i=1}^{N} \pi_i^{s_i}(1-\pi_i)^{1-s_i}, \qquad 0 < \pi_i < 1, s_i \in \{0,1\}. \end{aligned} \tag{13}$$

Weiterführend lässt sich diese Modellannahme unmittelbar mit der Vorstellung, dass $\vec{x}_i$ die Selektionswahrscheinlichkeit <u>vollständig</u> erklärt, in einen Zusammenhang bringen:

Mit $\pi_i := \Pr(S_i = 1 \mid \vec{x}_i)$ lässt sich die Wahrscheinlichkeit, dass i dem treatment zugewiesen wird, allgemein als eine, zumeist unbekannte, Funktion des individuellen Kovariablenvektors $\vec{x}_i$ auffassen (Rosenbaum, 1995), so dass:

$$\pi_i := \Pr(S_i = 1 \mid \vec{x}_i) = \lambda(\vec{x}_i) \qquad 0 < \lambda(\vec{x}_i) < 1.$$

Das vorherig formalisierte probabilistische und individualistische treatment-assignment (13) lässt sich entsprechend formulieren zu:

$$\begin{aligned}&\Pr(S_1 = s_1, \ldots, S_N = s_N \mid y_{t,1}, y_{c,1}, \vec{x}_1, \vec{z}_1, \ldots, y_{t,N}, y_{c,N}, \vec{x}_N, \vec{z}_N)\\ &= \Pr(S_1 = s_1, \ldots, S_N = s_N \mid \vec{x}_1, \ldots, \vec{x}_N)\\ &= \prod_{i=1}^{N} \lambda(\vec{x}_i)^{s_i}(1-\lambda(\vec{x}_i))^{1-s_i}. \end{aligned} \tag{14}$$

Unmittelbar ersichtlich wird die unterstellte Unabhängigkeit des Selektionsindikators von den potentiellen Werten, $y_{i,t}, y_{i,c}$, durch die Gleichsetzung der Wahrscheinlichkeiten:
$\Pr(S_i = s_i \mid y_{t,i}, y_{c,i}, \vec{x}_i, \vec{z}_i) = \Pr(S_i = s_i \mid \vec{x}_i)$; das Design ist bedingt an $\vec{x}_i$ dem Modell nach nicht mehr konfundiert, es gilt entsprechend:

$$(Y_t, Y_c) \perp\!\!\!\perp S_i \mid \vec{x}_i.$$

Zusätzlich formuliert $\Pr(S_i = s_i \mid y_{t,i}, y_{c,i}, \vec{x}_i, \vec{z}_i) = \Pr(S_i = s_i \mid \vec{x}_i)$ die Modellvorstellung, dass $\vec{z}_i$ unabhängig vom Selektionsindikator ist und eine Selektion nur in Abhängigkeit von $\vec{x}_i$ erfolgt, es resultiert empirisch ein overt bias, es liegt jedoch kein *hidden bias*, d.h. $F(\mathbf{Z} \mid S_i = 1) \neq F(\mathbf{Z} \mid S_i = 0)$, vor. Inhaltlich lässt sich der Kern dieses Selektionsmodells exemplarisch verdichten (Rosenbaum, 2010, S.70 f.):

> „(...) the (...) model would be true if treatments (...) were assigned by independent flips of a group of biased coins, where the same biased coin is used whenever

two people (...) have the same observed covariate (...) and no coin has probability 1 or 0 of a head. (...) it is not necessary that the biases of these coins be known, but it is necessary that the bias depend on $\mathbf{x}_l$ alone."

Das Selektionsmodell in (14) formuliert damit drei - empirisch nicht notwendigerweise prüfbare - Restriktionen an den Selektionsindikator S_i, die in Konsequenz eine an $\vec{x}_i$ bedingte Randomisierung implizieren:

1. Probabilistisches treatment-assignment: Die Restriktion, $0 < \Pr(S_i = 1 \mid \vec{x}_i) < 1$, formuliert die basale Annahme der Randomisierung: Gegeben dieser Vorstellung gilt, dass i konditional an $\vec{x}_i$ potentiell sowohl zu t als auch zu c zugewiesen werden kann. Die Gültigkeit dieser Modellannahme lässt sich empirisch prüfen, als dass bedingt an jedem $\vec{x}_i$ Einheiten dem treatment als auch der Kontrollbedingungen zugeordnet sind.

2. Individualistic treatment-assignment: formuliert eine Selektionwahrscheinlichkeit von i zum treatment, die unabhängig von der Selektion der anderen statistischen Einheiten ist. Für jedes i wird bedingt an $\vec{x}_i$ die Selektion durch einen „Münzwurf" vollzogen.

3. Unkonfundiertes treatment-assignment: formuliert die Restriktion, dass die an $\vec{x}_i$ bedingte Selektionswahrscheinlichkeit unabhängig von den potentiellen Werten sowie von nicht erhobenen Kovariablen $\vec{z}_i$ ist, d.h.:

$$\Pr(S_i = 1 \mid \vec{x}_i, \vec{z}_i, y_{i,t}, y_{i,c}) = \Pr(S_i = 1 \mid \vec{x}_i)$$

   Es handelt sich hierbei um eine nicht weiter prüfbare Annahme, da sich ein hidden bias, der einem treatment-assignment, das von $\vec{z}_i$ bestimmt wird, resultiert, nicht anhand von Daten beobachten lässt. Liegt empirisich ein overt bias vor, muss stets unterstellt werden, dass auch ein hidden bias vorliegt (Rosenbaum, 1995).

Das in (14) aufgestellte Selektionsmodell formuliert ein *strongly ignorable treatment-assignment*, das ein randomisiertes treatment-assignment bedingt an $\vec{x}_i$ beinhaltet und so bedingt an $\vec{x}_i$ eine Unabhängigkeit des Selektionsindikator von den potentiellen Variablen formuliert:

$$(Y_t, Y_c) \perp\!\!\!\perp S_i \mid \vec{x}_i \tag{15}$$

Die weitreichenden Konsequenzen werden folgend aufgewiesen.

1.8.3 Strongly Ignorable Treatment-Assignment

Es lässt sich gegeben der vorangestellten Modellannahme, unterhalb derer das an $\vec{x}_i$ bedingte treatment-assignment randomisiert erfolgt, zeigen, dass eine Identifikation und unkonfundierte, d.h. von $\vec{x}_i$ unabhängige Schätzung des ACE möglich ist. Kern aller folgender Überlegungen findet sich in der Modellannahme:

$$\Pr(S_i = 1 \mid y_{i,t}, y_{i,c}, \vec{x}_i, \vec{z}_i) = \Pr(S_i = 1 \mid \vec{x}_i), \quad 0 < \Pr(S_i = 1 \mid \vec{x}_i) < 1.$$

Diese Gleichsetzung formuliert damit eindeutig aus, dass gegeben $\vec{x}_i$ die Selektion randomisiert erfolgt und die potentiellen response-Variablen Y_t und Y_c sowie der ungemessene Kovariablenvektor $\vec{z}_i$ unabhängig vom Selektionsindikator S_i sind. Ein derartiges treatment-assignment wird als *strongly ignorable* klassifizert (z.B., Lunceford & Davidian, 2004; Rosenbaum, 1984b, 2010; Rubin, 1978; Rubin, Stuart & Zanutto, 2004; Stuart, 2010) - wie an späterer Stelle gezeigt wird, folgt diesem Modell für zwei Einheiten mit identischem $\vec{x}_i$, sie gelten in ihren Kovariablen als *exakt matched*, eine Chancengleichheit, zu t selegiert zu werden (z.B., Rosenbaum, 1995, 2010; Rubin, 1977). Demnach spielt gegeben einem strongly ignorable treatment-assignment die Selektion der Einheiten mit identischen $\vec{x}_i$ keine Rolle mehr und Einheiten, für die $S_i = 1$, liefern mit Beobachtung der zugehörigen potentiellen response-Werten diejenigen Informationen für den kontrafaktischen Erwartungswert $E(Y_t \mid S_i = 0, \vec{x}_i)$, die durch die Selektion verloren gegangen sind - die in $\vec{x}_i$ *matched* Gruppen sind homogen in ihren potentiellen Variablen, d.h.:

$$\begin{aligned} E(Y_t \mid S_i = 1, \vec{x}_i) &= E(Y_t \mid S_i = 0, \vec{x}_i) = E(Y_t \mid \vec{x}_i) \text{ sowie} \\ E(Y_c \mid S_i = 0, \vec{x}_i) &= E(Y_c \mid S_i = 1, \vec{x}_i) = E(Y_c \mid \vec{x}_i) \end{aligned}$$

In Konsequenz lässt sich gegeben einem *strongly ignorable treatment-assignment* der an $\vec{x}_i$ bedingte ACE durch Beobachtung entsprechend selegierter Gruppen identifizieren:

$$\begin{aligned} E(Y_t \mid S_i = 1, \vec{x}_i) - E(Y_c \mid S_i = 0, \vec{x}_i) &= E(Y_t \mid S_i = 1, \vec{x}_i) - E(Y_c \mid S_i = 1, \vec{x}_i) \\ &= E(Y_t \mid \vec{x}_i) - E(Y_c \mid \vec{x}_i) \\ &= E(Y_t - Y_c \mid \vec{x}_i) \end{aligned} \tag{16}$$

Dem folgend lässt sich gemäß dem *law of iterated expectations* der unkonditionierte ACE über alle $\vec{x}_i$ hinweg identifizieren, es gilt:

$$\begin{aligned} E_{\vec{x}_i}\{E(Y_t \mid S_i = 1, \vec{x}_i) - E(Y_c \mid S_i = 0, \vec{x}_i)\} &= E_{\vec{x}_i}\{E(Y_t \mid \vec{x}_i) - E(Y_c \mid \vec{x}_i)\} \\ &= E_{\vec{x}_i}\{E(Y_t - Y_c \mid \vec{x}_i)\} \\ &= E(Y_t - Y_c) \\ &= E(Y_t) - E(Y_c) \end{aligned} \tag{17}$$

Damit lässt sich unter den Annahmen, dass

a) $\vec{x}_i$ sämtliche Kovariablenwerte, die die treatment-Selektion bedingen, enthält und

b) gegeben $\vec{x}_i$ eine positive Wahrscheinlichkeit, in das treatment selegiert zu werden, besteht,

formell begründen, dass in observational studies unter der Annahme eines strongly ignorable treatment-assignment der ACE identifiziert werden kann. Rubin (1977, S.2) formuliert bezüglich einer möglichen Schätzung des ACE in Stichproben inhaltlich aus:

> „(...) the appropiate estimate of τ is the average value of the difference between the estimated conditional expectation of Y on X in the two treatment groups.“

Als Schätzstatistik für den an $\vec{x}_i$ bedingten ACE lässt sich damit die an $\vec{x}_i$ bedingte Mittelwertsdifferenz definieren:

$$\bar{Y}_{t|\vec{x}_i} - \bar{Y}_{c|\vec{x}_i} = \frac{1}{n_t} \sum_{i:i\in\{I_i=1,S_i=1,\vec{x}_i\}} y_i - \frac{1}{n_c} \sum_{i:i\in\{I_i=1,S_i=0,\vec{x}_i\}} y_i \tag{18}$$

Die Erwartungstreue dieser Schätzstatistik lässt sich den gesamten Ausführungen nach entsprechend zeigen:

$$\begin{aligned} E\left(\bar{Y}_{t|\vec{x}_i} - \bar{Y}_{c|\vec{x}_i}\right) &= E(Y_t \mid S_i = 1, \vec{x}_i) - E(Y_c \mid S_i = 0, \vec{x}_i) \\ &= E(Y_t \mid \vec{x}_i) - E(Y_c \mid \vec{x}_i). \end{aligned}$$

Gemäß dem Fall, dass $\vec{x}_i$ diskret und nicht hochdimensional ist folgt damit als mögliche Schätzstatistik für den unkonditionierten ACE (z.B., Emura, Wang & Katsuyama, 2008; Rubin, 1977):

$$\hat{\text{ACE}} = \sum_{\vec{x}_i} \frac{n_{\vec{x}_i}}{n} \left(\frac{1}{n_t} \sum_{i:i\in\{I_i=1,S_i=1,\vec{x}_i\}} y_i - \frac{1}{n_c} \sum_{i:i\in\{I_i=1,S_i=0,\vec{x}_i\}} y_i \right), \tag{19}$$

mit $n_{\vec{x}_i}$: Anzahl der $n-$Einheiten mit $\vec{x}_i$, d.h.: $n_{\vec{x}_i} =\mid \{i \mid I_i = 1 \cap \vec{x}_i\} \mid$.

Entscheidend damit: Bei Gültigkeit der aufgewiesenen Modellannahmen lässt sich eine von $\vec{x}_i$ unabhängige Schätzung des ACE ermöglichen, so dass entsprechende Störeinflüsse der Kovariablen behoben werden können. Es handelt sich bei einer solchen Schätzung um ein Vorgehen, das folgend als *Adjustierung* definiert ist.

1.9 Propensity Score

1.9.1 Definition: Balancing Score

Sämtliche vorangegangene Überlegungen zu dem Selektionsmodell in observational studies führten einerseits zu einer an $\vec{x}_i$ bedingten Randomisierung, die eine Identifikation des ACE in observational studies ermöglichte, andererseits zu einem overt bias, der sich in den Daten mit $F(\mathbf{X} \mid S_i = 1) \neq F(\mathbf{X} \mid S_i = 0)$ dem Selektionsmodell folgend niederschlägt und eine mögliche Konfundierung durch die erhobenen Kovariablen offenlegt. Demnach erscheint es naheliegend, in einer observational study bei vorliegenden overt bias eine *Adjustierung* der entsprechenden Kovariablen bei der Schätzung des ACE vorzunehmen, so dass in Konsequenz beide Experimentalgruppen hinsichtlich dieser Kovariablen, die entsprechende Störvariablen darstellen könnten, homogen verteilt sind; so, wie es einer Randomisierung gefolgt wäre. Im Folgenden wird damit der Begriff der *Adjustierung* gleichgesetzt mit derjenigen Methodik, durch die ein overt bias behoben werden kann und der folgend der ACE unverzerrt geschätzt werden kann (Rosenbaum, 1995). In diesem Zusammenhang führen Rosenbaum und Rubin (1983) den sog. *balancing score*, $b(\vec{x}_i)$, ein: $b(\vec{x}_i)$ als eine allgemeine Funktion des Kovariablenvektors $\vec{x}_i$ ist grundlegend dergestalt, dass konditional an $b(\vec{x}_i)$ die konfundierenden Kovariablen über beide Experimentalgruppen identisch verteilt sind, so dass:

$$F(\mathbf{X} \mid S_i = 1, b(\vec{x}_i)) = F(\mathbf{X} \mid S_i = 0, b(\vec{x}_i)).$$

Der identischen Verteilung der Kovariablen folgend können diese bedingt an $b(\vec{x}_i)$ keine Störvariablen darstellen und die Verteilung der beobachteten Kovariablen entspricht der bei Randomisierung. Der plausibelste balancing score ist $b(\vec{x}_i) = \vec{x}_i$ - ein Fall der folgend als *exaktes matching* definiert und an späterer Stelle expliziert wird.

Des weiteren ist $b(\vec{x}_i)$ als eine Funktion der Kovariablen so zu wählen, dass bedingt an $b(\vec{x}_i)$ ein strongly ignorable treatment-assignment hält; konditional an $b(\vec{x}_i)$ müssen damit folgende

Annahmen erfüllt werden (vgl. 1.8.2):

a) $\mathbf{X} \perp\!\!\!\perp S_i \mid b(\vec{x}_i)$

b) $(Y_t, Y_c) \perp\!\!\!\perp S_i \mid b(\vec{x}_i)$

c) $0 < b(\vec{x}_i) < 1$.

Unter Gültigkeit dieser Annahmen nach unterscheiden sich die statistischen Einheiten konditional an $b(\vec{x}_i)$ nur noch in ihrer Zugehörigkeit zu einer der Versuchsbedingungen, wobei gegeben $b(\vec{x}_i)$ eine randomisierte Selektion zugrunde gelegt werden kann. Damit ist das Design konditional an $b(\vec{x}_i)$ unkonfundiert und es liegen bedingt an $b(\vec{x}_i)$ Gruppen vor, die in jeglichen Variablen homogen sind, wie es einem randomisierten treatment-assignment resultiert wäre; es gilt damit:

$$E(Y_t \mid S_i = 1, \vec{x}_i, b(\vec{x}_i)) = E(Y_t \mid S_i = 0, \vec{x}_i, b(\vec{x}_i)) = E(Y_t \mid \vec{x}_i, b(\vec{x}_i))$$

sowie

$$E(Y_c \mid S_i = 0, \vec{x}_i, b(\vec{x}_i)) = E(Y_c \mid S_i = 1, \vec{x}_i, b(\vec{x}_i)) = E(Y_c \mid \vec{x}_i, b(\vec{x}_i)),$$

die Gruppen würden demnach entsprechende Informationen über die kontrafaktischen Erwartungswerte liefern und der ACE wäre konditional an $b(\vec{x}_i)$ identifizierbar.

1.9.2 Definition: Propensity Score $e(\vec{x}_i)$

Von besonderem Interesse als balancing score ist der sog. *propensity score* der Einheit i, folgend $e(\vec{x}_i)$; $e(\vec{x}_i)$ ist definiert als die bedingte Wahrscheinlichkeit für i gegeben $\vec{x}_i$ ins treatment zu gelangen:

$$e(\vec{x}_i) := \Pr(S_i = 1 \mid \vec{x}_i), \quad 0 < e(\vec{x}_i) < 1. \tag{20}$$

Gemäß dem Satz von Bayes folgt mit $s_i \in \{0, 1\}$:

$$e(\vec{x}_i) = \frac{\Pr(\vec{x}_i \mid S_i = 1) \cdot \Pr(S_i = 1)}{\Pr(\vec{x}_i \mid S_i = 1) \cdot \Pr(S_i = 1) + \Pr(\vec{x}_i \mid S_i = 0) \cdot \Pr(S_i = 0)}.$$

Gegeben dem Satz von Bayes wird ersichtlich, dass der propensity score von i sämtliche Informationen über das angenommene treatment-assignment enthält: $\Pr(\vec{x}_i \mid S_i = 1)$ berücksichtigt die gemeinsame Verteilung von $\mathbf{X}$ und S_i, die das treatment-assignment definiert; entsprechend

lässt das treatment-assignment in observational studies - aufgeführt in (14) - mit Definition des propensity scores formalisieren zu:

$$\begin{aligned}\Pr(S_1 = s_1, ..., S_N = s_N \mid \vec{x}_1, ..., \vec{x}_N) &= \prod_{i=1}^{N} \lambda(\vec{x}_i)^{s_i}(1 - \lambda(\vec{x}_i))^{1-s_i} \\ &= \prod_{i=1}^{N} e(\vec{x}_i)^{s_i}(1 - e(\vec{x}_i))^{1-s_i} \qquad (21)\end{aligned}$$

Im Unterschied zu vollständig randomisierten Experimenten, in denen der propensity score von i bekannt ist - es gilt: $e(\vec{x}_i) = \Pr(S_i = 1 \mid \vec{x}_i) = \Pr(S_i = 1)$ - ist in observational studies $e(\vec{x}_i)$ unbekannt, da definitionsgemäß mit Datenerhebung die Zuordnungen zu den Experimentalbedingungen vorliegen. Folglich ist unbekannt, welche Kovariablen die Selektion bedingen und es muss zu Beginn einer observational study bei Vorliegen eines overt bias $e(\vec{x}_i)$, z.B. durch eine logistische Regression oder eine Diskriminanzanalyse, gegeben der beobachteten Daten $(\vec{x}_i, s_i)$ geschätzt werden, wobei entsprechend vorausgesetzt werden muss, dass $\vec{x}_i$ all diejenigen Werte enthält, die i zur Wahl des treatments motivieren.

1.10 $e(\vec{x}_i)$ als Balancing Score

1.10.1 Theoreme und Beweise

Folgend wird aufgewiesen, dass $e(\vec{x}_i)$ der *coarest*[5] balancing score ist, gegeben dem ein strongly ignorable treatment-assignment hält (z.B., Rosenbaum, 1995; Rosenbaum & Rubin, 1983):

Theorem 1. *Der propensity score $e(\vec{x}_i)$ ist ein balancing score, formal:*

$$\mathbf{X} \perp\!\!\!\perp S_i \mid e(\vec{x}_i).$$

Demnach liegt bei Gültigkeit dieses Theorems bedingt an $e(\vec{x}_i)$ eine Unabhängigkeit von $\vec{x}_i$ und S_i vor, so dass:

$$\Pr(\vec{x}_i, S_i = 1 \mid e(\vec{x}_i)) = \Pr(\vec{x}_i \mid e(\vec{x}_i)) \cdot \Pr(S_i = 1 \mid e(\vec{x}_i))$$

Grundsätzlich gilt für drei Ereignisse:

$$\Pr(\vec{x}_i, S_i = 1 \mid e(\vec{x}_i)) = \Pr(\vec{x}_i \mid e(\vec{x}_i)) \cdot \Pr(S_i = 1 \mid \vec{x}_i, e(\vec{x}_i)),$$

[5] d.h., der *grobmaschigste*

da $e(\vec{x}_i)$ eine Funktion von $\vec{x}_i$ ist, gilt vereinfacht:

$$\Pr(S_i = 1 \mid \vec{x}_i, e(\vec{x}_i)) = \Pr(S_i = 1 \mid \vec{x}_i).$$

Damit genügt es folgend zu zeigen, dass

$$\Pr(S_i = 1 \mid \vec{x}_i) = \Pr(S_i = 1 \mid e(\vec{x}_i)).$$

Beweis. Per definitionem gilt:

$$\Pr(S_i = 1 \mid \vec{x}_i) = e(\vec{x}_i).$$

Mit $s_i \in \{0, 1\}$ folgt unter Nutzung des *law of iterated expectations*:

$$\Pr(S_i = 1 \mid e(\vec{x}_i)) = E(S_i \mid e(\vec{x}_i)) = E\{E(S_i \mid \vec{x}_i) \mid e(\vec{x}_i)\} = E(e(\vec{x}_i) \mid e(\vec{x}_i)) = e(\vec{x}_i)$$

Damit:

$$\Pr(S_i = 1 \mid \vec{x}_i) = \Pr(S_i = 1 \mid e(\vec{x}_i))$$

□

Theorem 2. *Sei $b(\vec{x}_i)$ eine Funktion von $\vec{x}_i$; $b(\vec{x}_i)$ ist nur dann ein balancing score, wenn er feiner als der propensity score $e(\vec{x}_i)$ ist, so dass $e(\vec{x}_i) = f(b(\vec{x}_i))$, mit einer beliebigen Funktion $f(\cdot)$, und:*

$$\mathbf{X} \perp\!\!\!\perp S_i \mid b(\vec{x}_i),$$

Beweis. Insofern $b(\vec{x}_i)$ ein balancing score ist, so genügt es zu zeigen, dass:

$$\Pr(S_i = 1 \mid b(\vec{x}_i)) = e(\vec{x}_i)$$

Da:

$$e(\vec{x}_i) = \Pr(S_i = 1 \mid \vec{x}_i) = E(S_i \mid \vec{x}_i),$$

muss entsprechend mit $e(\vec{x}_i) = f(b(\vec{x}_i))$ gelten:

$$\Pr(S_i = 1 \mid b(\vec{x}_i)) = E\{E(S_i \mid \vec{x}_i) \mid b(\vec{x}_i)\}.$$

Es folgt aus der Definition $e(\vec{x}_i) = \Pr(S_i = 1 \mid \vec{x}_i)$ und dem *law of iterated expectations*:

$$E\{E(S_i \mid \vec{x}_i) \mid b(\vec{x}_i)\} = E\{e(x_i) \mid b(\vec{x}_i)\} = e(x_i)$$

□

Theorem 3. *Wenn bedingt an $\vec{x}_i$ das treatment-assignment strongly ignorable ist, d.h.:*

$$(Y_t, Y_c) \perp\!\!\!\perp S_i \mid \vec{x}_i \quad \textit{und} \quad 0 < \Pr(S_i = 1 \mid \vec{x}_i) < 1,$$

so ist konditional an $b(\vec{x}_i)$ das treatment-assignment strongly ignorable, d.h.:

$$(Y_t, Y_c) \perp\!\!\!\perp S_i \mid b(\vec{x}_i) \quad \textit{und} \quad 0 < \Pr(S_i = 1 \mid b(\vec{x}_i)) < 1$$

Beweis. Es genügt zu zeigen, dass:

$$\Pr(S_i = 1 \mid y_{i,t}, y_{i,c}, b(\vec{x}_i)) = \Pr(S_i = 1 \mid b(\vec{x}_i)) = e(x_i)$$

Da:

$$\Pr(S_i = 1 \mid y_{i,t}, y_{i,c}, b(\vec{x}_i)) = E\{\Pr(S_i = 1 \mid y_{i,t}, y_{i,c}, \vec{x}_i) \mid y_{i,t}, y_{i,c}, b(\vec{x}_i)\}$$

und $(Y_t, Y_c) \perp\!\!\!\perp S_i \mid \vec{x}_i$, folgt:

$$E\{\Pr(S_i = 1 \mid \vec{x}_i) \mid y_{i,t}, y_{i,c}, b(\vec{x}_i)\} = E\{e(x_i) \mid y_{i,t}, y_{i,c}, b(\vec{x}_i)\} = e(x_i)$$

□

Theorem 4. *Den Theoremen 1-3 folgt, dass bedingt an $b(\vec{x}_i)$ die Differenz der an S_i bedingten Erwartungswerte der potentiellen response-Variablen dem ACE konditional an $b(\vec{x}_i)$ entspricht, d.h.:*

$$E(Y_t \mid S_i = 1, b(\vec{x}_i)) - E(Y_c \mid S_i = 0, b(\vec{x}_i)) = E(Y_t \mid b(\vec{x}_i)) - E(Y_c \mid b(\vec{x}_i))$$

Beweis. Aus Theorem 3 folgt:

$$E_{b(\vec{x}_i)}\{E(Y_t \mid S_i = 1, b(\vec{x}_i)) - E(Y_c \mid S_i = 0, b(\vec{x}_i))\}$$

$$= E_{b(\vec{x}_i)}\{E(Y_t \mid b(\vec{x}_i)) - E(Y_c \mid b(\vec{x}_i))\} = E(Y_t) - E(Y_c)$$

□

In Dawid's Notation lassen sich damit vorherige Überlegungen verdichten zu:

$$\mathbf{X} \perp\!\!\!\perp S_i \mid e(\vec{x}_i) \text{ sowie } (Y_t, Y_c) \perp\!\!\!\perp S_i \mid e(\vec{x}_i), \tag{22}$$

so dass konditional an $e(\vec{x}_i)$ ein strongly ignorable treatment-assignment, genauso wie an $\vec{x}_i$, hält. Rosenbaum und Rubin (1983, S. 44) weisen auf die praktische Implikation dieser bedingten Unabhängigkeit hin:

„(...) if a subclass of units or a matched treatment-control pair is homogeneous in $e(x)$, then the treated and control units in that subclass or matched pair will have the same distribution of x.“

Durch Konditionierung an $e(\vec{x}_i)$ als balancing score ergibt sich über die Experimentalbedingungen hinweg eine identische Verteilung der Kovariablen und potentiellen Variablen; die statistischen Einheiten unterscheiden sich damit nur noch in ihrem Selektionsstatus, der gegeben einem strongly ignorable treatment-assignment als an $\vec{x}_i$ bedingt randomisiert zustande gekommen angenommen werden kann. Den Theoremen nach hält eine an $e(\vec{x}_i)$ bedingte Randomisierung, d.h., dass gegeben $e(\vec{x}_i)$ das treatment-assignment ebenfalls strongly ignorable ist, so dass die potentiellen response-Variablen bedingt an $e(\vec{x}_i)$ unabhängig vom Selektionsindikator sind und der ACE bedingt an $e(\vec{x}_i)$ identifizierbar ist, d.h.:

$$\begin{aligned} & E_{e(\vec{x}_i)}\{E(Y_t \mid S_i = 1, e(\vec{x}_i)) - E(Y_c \mid S_i = 0, e(\vec{x}_i))\} \\ = & E_{e(\vec{x}_i)}\{E(Y_t \mid e(\vec{x}_i)) - E(Y_c \mid e(\vec{x}_i))\} = E(Y_t) - E(Y_c) \end{aligned} \tag{23}$$

Rosenbaum und Rubin (1983, S.46) formulieren entsprechend aus:

„In words, under strongly ignorable treatment assignment, units with the same value of the balancing score $b(x)$ but different treatments can act as controls for each other, in the sense that the expected difference in their response equals the average treatment effect.“

1.10.2 Praktische Implikationen

Sämtliche Theoreme, Beweise und daraus abgeleiteten Implikationen beziehen sich auf den theoretischen propensity score $e(\vec{x}_i)$, der in einer observational study de facto nicht bekannt ist und anhand der Stichprobendaten $(s_i, \vec{x}_i)$ geschätzt werden muss.

Es lässt sich jedoch deduzieren: Eine Adjustierung mit einem in Stichproben geschätzten propensity score, $\hat{e}(\vec{x}_i)$, führt grundlegend immer zu einer Balance derjenigen Kovariablen, die bei der Modellierung des propensity scores benutzt worden sind (z.B., Rosenbaum, 1987; Rosenbaum & Rubin, 1983; Stuart, 2010), unabhängig davon, ob das zu Grunde gelegte Modell dem treatment-assignment entspricht, so dass die Funktion als balancing score stets garantiert ist.

Des Weiteren gilt: Einer vollständigen und korrekten Spezifikation des propensity scores-Modells in Stichproben, so dass $\hat{e}(\vec{x}_i)$ ein konsistenter Schätzer für $e(\vec{x}_i)$ ist, folgt ein strongly ignorable treatment-assigment, so dass bedingt an $\hat{e}(\vec{x}_i)$ der ACE identifiziert werden und durch Methoden, basierend auf dem theoretischen propensity score, $e(\vec{x}_i)$, konsistent geschätzt werden kann (s. hierzu Kapitel 1.11).

Es muss an dieser Stelle jedoch betont werden, dass $e(\vec{x}_i)$ - auch bei einer korrekten und vollständigen Spezifikation des Selektionsmodells, die sämtliche Kovariablen, Interaktionen der Kovariablen und höhere Polynome beinhaltet - eine Balance der Verteilungen der beobachteten Kovariablen $\mathbf{X}$ garantiert; für nicht erhobene Kovariablen $\mathbf{Z}$ kann dies nicht sichergestellt werden, wie Rosenbaum und Rubin (1984, S.517) betonen:

> „Of course, although we expect subclassification on an estimated $e(\mathbf{x})$ to produce balanced distributions of $\mathbf{x}$, it cannot, like randomization, balance unobserved covariates, except to the extent that they are correlated with $\mathbf{x}$. “

Der entscheidende Vorteil bei der Nutzung des propensity scores soll kurz skizziert werden: Wie in (17) und (18) aufgewiesen, ist bedingt an $\vec{x}_i$ einerseits eine Identifikation des ACE, andererseits eine erwartungstreue und konsistente Schätzung des ACE möglich. Da jedoch mit zunehmender Dimensionalität von $\vec{x}_i$ die Anzahl an möglichen Kovariablenvektoren steigt, wird ein direktes matching, wie dem Modell nach impliziert wird, zunehmend unwahrscheinlicher. Mit p binären Kovariablen lassen sich entsprechend 2^p mögliche Vektoren $\vec{x}_i$ finden, so dass die Wahrscheinlichkeit für einen exakten match mit zunehmenden p gegen null konvergiert. Da $e(\vec{x}_i)$ als bedingte Wahrscheinlichkeit skalar ist, lässt sich ein multidimensionales matching durch ein *propensity score*-matching ersetzen. Dies wird an späterer Stelle ausgeführt.

1.10.3 Schätzung des Propensity Scores

Das binomiale logistische Regressionsmodell

In praxi ist der propensity score $e(\vec{x}_i)$ ein unbekannter Wert, der in einer observational study ausgehend von den Stichprobenrealisierungen $(s_i, \vec{x}_i)$ geschätzt werden muss. Dazu wird in praxi zumeist das parametrische Verfahren der logistischen Regression gewählt (z.B., Austin, 2011; Rosenbaum, 1984b; Rosenbaum & Rubin, 1984; Stuart, 2010), welches folgend kurz dargestellt werden soll.

Ziel des logistischen Regressionsmodells ist die Schätzung der Effekte von $p-$Kovariablen mit den Werten $\vec{x}_i^t = (x_{i1}, ..., x_{ip})$ auf die bedingte Wahrscheinlichkeit

$$\pi_i = \Pr(Y_i = 1 \mid x_{i1}, ..., x_{ip}) = E(Y_i \mid \vec{x}_i),$$

wobei für folgende Zwecke der Regressant Y_i durch den Selektionsindikator S_i, mit $s_i \in \{0, 1\}$, ersetzt werden kann, so dass:

$$\pi_i = \Pr(S_i = 1 \mid \vec{x}_i) = E(S_i \mid \vec{x}_i) = e(\vec{x}_i).$$

Die entscheidende Erweiterung, die das logistische Regressionsmodell im Verhältnis zum klassischen Regressionsmodell bietet: die Werte der abhängigen Variablen Y_i dürfen binär kodiert sein und müssen an der Stelle nicht einem mindestens intervallskalierten, stetigen Merkmal entsprechen. Damit wird an den linearen Prädiktor des logistischen Regressionsmodell, η_i, wobei η_i als die systematische Komponente des Regressionsmodells, welche die lineare Effekte der Kovariablen modelliert, definiert ist, die Restriktion aufgestellt, dass dieser nur Werte im Intervall $[0, 1]$ liefert. Damit ist η_i formal definiert als

$$\eta_i = \beta_0 + \beta_1 \cdot x_{i1} + ... + \beta_p \cdot x_{ip} = \vec{x}_i^t \vec{\beta}$$

Der Restriktion $\eta_i \in [0, 1]$ wird Genüge geleistet durch die Verbindung der bedingten Wahrscheinlichkeit π_i mit dem linearen Prädiktor η_i über die *response-Funktion* $h(\eta_i) \in [0, 1]$, die im Falle des logistischen Regressionsmodells eine streng monoton wachsende Verteilungsfunktion auf der gesamten reellen Achse darstellt. Die Verbindung von π_i zu η_i ist definiert durch

$$\pi_i = h(\eta_i) = h(\vec{x}_i^t \vec{\beta}).$$

Die Umkehrung erfolgt durch die *link-Funktion:*

$$\eta_i = g(\pi_i) \quad \text{wobei } g = h^{-1}$$

Im Logit-Modell, als Spezialfall des logistischen Regressionsmodell, welches zumeist zu Grunde gelegt wird, ist $h(\eta_i)$ definiert als logistische response-Funktion mit:

$$\pi_i = \frac{\exp(\eta_i)}{1 + \exp(\eta_i)},$$

die link-Funktion $g(\pi_i)$ als:

$$g(\pi_i) = \log\left(\frac{\pi_i}{1 - \pi_i}\right) = \eta_i = \vec{x}_i^t \vec{\beta}.$$

Die Interpretation der Effekte der Regressoren auf den Regressanten (s. link-Funktion) sind *log-odd's*, die nach Transformation überführt werden können in entsprechende Chancen-Verhältnisse, d.h.:

$$\begin{aligned}\frac{\pi}{1-\pi} &= \exp(\beta_0) \cdot \exp(\beta_1 x_1) \cdot ... \cdot \exp(\beta_p x_p) \\ &= \frac{\Pr(y_i = 1 \mid \vec{x}_i)}{\Pr(y_i = 0 \mid \vec{x}_i)}\end{aligned}$$

Im Falle des propensity scores lassen sich die geschätzten Effekte interpretieren als

$$\frac{\Pr(S_i = 1 \mid \vec{x}_i)}{\Pr(S_i = 0 \mid \vec{x}_i)},$$

demnach als Chance, gegeben einem Kovariablenvektor $\vec{x}_i$ ins treatment bzw. in die Kontrollbedingung zu gelangen. Durch einfache Umformung lassen sich aus den entsprechenden Chancen die Selektionswahrscheinlichkeiten bestimmen, denn es gilt:

$$\exp(g(\pi)) = \frac{\pi}{1-\pi} \leftrightarrow \pi = \frac{\exp(g(\pi))}{(1+\exp(g(\pi)))}$$

***Boosted logistic regression* zur Schätzung von $e(\vec{x}_i)$**

McCaffrey, Ridgeway und Morral (2004) weisen in ihrem Artikel auf den Umstand hin, dass eine konsistente Schätzung des propensity scores, $e(\vec{x}_i)$, durch das logistische Regressionsmodell voraussetzt, dass die in das Modell eingeschlossenen Prädiktoren linear und additiv auf die log-odd-Skala zur Modellierung der Selektionswahrscheinlichkeit einwirken und mögliche Interaktionen zwischen den Prädiktoren oder nicht-lineare Terme entsprechend modelliert werden müssen. Zumeist liegt in praxi keine Kenntnis über den funktionalen Zusammenhang der Kovariablen und dem treatment-assignment vor. Die parametrische Modellierung des propensity scores durch das logistische Regressionsmodell beinhaltet demnach immer das Risiko, dass relevante Kovariablen nicht in das Modell aufgenommen werden (sog. *ommited variables*-Problem (z.B., Heckman, 1979; Winship & Mare, 1992)) oder der Zusammenhang zwischen einer Kovariablen und der treatment-Selektion nicht richtig funktional spezifiziert ist. In einem viel zitierten Artikel weist Drake (1993) in mehreren Simulationsstudien die Konsequenzen eines fehlspezifizierten propensity scores-Modells auf: Insofern ein falsch modellierter propensity score zur Adjustierung genutzt wird, resultiert empirisch ein bias bei der Schätzung des ACE, der in bestimmten Fällen größer sein kann als der einer unadjustierten Schätzung - nach Stuart

(2010) ein *Paradoxon*, da entsprechend der propensity score genutzt wird, um diesen bias zu verringern.

Ausgehend von der Problematik, die sich bei der Spezifikation des treatment-assignments zur Schätzung des propensity scores ergibt, schlugen McCaffrey et al. (2004) vor, $e(\vec{x}_i)$ bzw., aus numerischer Einfachheit heraus, das zugehörige log-odd's ratio $g(\vec{x}_i) = \log(e(\vec{x}_i)/(1-e(\vec{x}_i)))$, ausgehend von beobachteten Kovariablen über ein *generalized boosted model* (GBM) zu schätzen. Die Autoren führen aus:

> „GBM is a general, automated, data-adaptive modeling algorithm that can estimate the nonlinear relationship between a variable of interest and a large number of covariates. (...) it can predict treatment assignment from a large number of pretreatment covariates while also allowing for flexible, nonlinear relationships between the covariates and the propensity score.“ (McCaffrey et al., 2004, S.404)

Das GBM versteht sich <u>nicht</u> als ein non-parametrisches Verfahren, den unbekannten Parametervektor $\vec{\beta}$ aus einem gegebenen Datensatz für die logisitische Regression zur Modellierung der Selektionswahrscheinlichkeit schätzt, sondern als ein non-parametrisches *Prädiktionsverfahren*, welches basierend auf einer Vielzahl von Kovariablen versucht, die an $\vec{x}_i$ bedingte Selektionswahrscheinlichkeit im Sinne eines minimalen Prädiktionsfehlers optimal zu schätzen, wobei zur Spezifikation des funktionalen Zusammenhangs von $\vec{x}_i$ und der Selektionswahrscheinlichkeit der *regresson tree*-Ansatz gewählt wird (Guo & Fraser, 2010).

Präziser wird durch die Anwendung des GBM versucht, eine *initiale* Schätzung für das log-odd's ratio, $\hat{g}(\vec{x}_i) = \log(\bar{S}_i/(1-\bar{S}_i))$, durch Hinzunahme von Funktionen der Kovariablenwerte, $h(\vec{x}_i)$, in dem Sinne optimieren zu können, als dass der resultierende Modell-Fit (sog. *Bernoulli log-likelihood*) maximiert wird.

> „GBMs add together many simple functions to estimate a smooth function of a large number of covariates. Each individual simple function lacks smoothness and is a poor approximation to the function of interest (...). In our implementation of GBMs, each simple function is a regression tree with limited depth.“
>
> (McCaffrey et al., 2004, S. 407)

Als Modell-Fit, der die Güte der getätigten Prädiktion quantifiziert, wird die *Bernoulli log-likelihood*-Funktion gewählt, die in Hinblick auf einen minimalen Prädiktionsfehler maximiert

werden soll, wobei als Prädiktionsfehler die Differenz vom treatment-Status (s_i) zu dem geschätzten propensity score, d.h.: $\hat{\epsilon}_i = s_i - \hat{e}(\vec{x}_i)$, gewählt wird. Die Bernoulli log-likelihood ist definiert als:

$$l\left(g(\vec{x}_i)\right) = \log(\mathcal{L}(g(\vec{x}_i))) = \sum_{i=1}^{n} s_i \cdot g(\vec{x}_i) - \log\left(1 + \exp[g(\vec{x}_i)]\right), \tag{24}$$

wobei im ersten Schritt $g(\vec{x}_i) = \hat{g}(\vec{x}_i)$, wie oben definiert, gewählt wird. Im nächsten Schritt wird eine Adjustierung der Form $h(\vec{x}_i)$ gesucht, die folgende Eigenschaft erfüllt:

$$l(g(\vec{x}_i) + \lambda \cdot h(\vec{x}_i)) > l(g(\vec{x}_i)) \qquad 0 < \lambda < 1.$$

Insofern ein $h(\vec{x}_i)$als Ergebnis eines möglichen *regression trees* gefunden werden kann, welches die resultierende Bernoulli log-likelihood-Funktion maximiert, so resultiert zur Schätzung des log-odd's ratio: $g(\vec{x}_i) = \hat{g}_(\vec{x}_i) + \lambda \cdot h(\vec{x}_i)$. Im vorliegenden Vorschlag zur Prädiktion des propensity scores wählten die Autoren $h(\vec{x}_i)$ als Schätzergebnis aus einem der möglichen *regression trees*. Die Idee des *regression trees* als Algorithmus zur Ermittlung von $h(\vec{x}_i)$ soll folgend skizziert werden:

Es wird der gesamte Datensatz auf Basis einer Kovariable aufgeteilt in zwei sog. *splits*, wobei eine solche Aufteilung bei jedem beliebigen Paar an Werten einer beliebigen Kovariable auftreten kann und an die Messdignität der Kovariablen keinerlei Anforderungen gestellt werden. Separat für jeden resultierenden tree wird geprüft, ob die Bernoulli log-likelihood durch Hinzunahme des trees bei der Prädiktion des propensity scores maximiert wird. Es handelt sich demnach bei der Ermittlung der regression trees um einen rekursiven Algorithmus; McCaffrey et al. (2004, S. 407) heben hervor:

> „Among all the possible splits, the algorithm selects the one that minimizes prediction error.“

Entscheidend: Nach der Schätzung eines $h(\vec{x}_i)$ wird dieses in die Funktion $g(\vec{x}_i)$ aufgenommen, so dass diese zunächst die Form $g(\vec{x}_i) = \hat{g}(\vec{x}_i) + \lambda \cdot h(\vec{x}_i)$ besitzt. Dieses Verfahren ist insofern als rekursiv zu verstehen, als dass ein $h(\vec{x}_i)$ nicht mit in die Gleichung $g(\vec{x}_i)$ genommen wird, wenn es die log-likelihood nicht maximiert; an dieser Stelle werden weitere splits ermittelt, bis das Minimierungskriterium der Prädiktionsfehler erfüllt ist.

Im Anschluss an die Auswahl des ersten splits, welches dem Kriterium des minimalen Prädiktionsfehler genügt, werden weitere Aufteilungen und Schätzungen des propensity scores in

jedem split vorgenommen, bis die vorher festgelegte Anzahl an maximalen splits erreicht ist. Der Vorteil bei diesem Vorgehen ist die automatische Ermittlung von möglichen Interaktionen, die sich durch einen split einer Kovariablen mit einem sequentiell folgenden split einer anderen Kovariable ergibt, wobei die Aufnahme als Interaktion in das Modell dann erfolgt, wenn an dieser Stelle ein Minimum des Vorhersagefehlers erreicht ist.

Weitere Vorschläge zur Schätzung des Propensity Scores

Schwerpunktmäßig konzentrieren sich neuere Veröffentlichungen im Themenbereich der *kausalen Inferenz* auf die konsistente Schätzung des unbekannten propensity scores, die dem folgend verwandt wird, um eine adjustierte, konsistente Schätzung des ACE zu ermöglichen. Folgend sollen kurz weitere Methoden zur Schätzung des unbekannten propensity scores in Stichproben vorgestellt werden:

Ausgehend von dem mehrfach aufgeführten Befund (z.B., Hahn, 1998; Heckman et al., 1997), dass eine Adjustierung an einem korrekt spezifizierten propensity score eine konsistente Schätzung des ACE ermöglicht, jedoch diese Schätzungen in Stichproben eine größere Varianz aufweisen als eine direkte Adjustierung gegeben den konfundierenden Kovariablen, schlugen Hirano, Imbens und Ridder (2003) einen asymptotisch effizienten, non-parametrischen Schätzer für den unbekannten propensity score vor, der für ein nachträgliches Gewichten der Beobachtungen bei der Schätzung des ACE genutzt werden und in Konsequenz zu einer asymptotisch effizienten Schätzung des ACE führt. Analog schlägt auch Hahn (1998) einen an dem propensity score adjustierten, non-parametrischen Schätzer für den ACE bzw. den in (9) definierten ATT vor, der asymptotisch effizient ist. Beide Vorschläge zeichnen sich dadurch aus, dass die Varianz der Schätzer *bounded* ist.

Imai und Ratkovic (2014) definieren in ihrem Artikel die *covariate balancing propensity score*-Methodologie; Ziel ist es, eine Schätzung des propensity scores zu ermöglichen, welche einerseits das zu Grunde liegende treatment-assignment modelliert, andererseits eine Balance der Kovariablen ersucht, um so bedingte Unabhängigkeit der Kovariablen vom Selektionsindikator plausibel zu machen; die getätigte Schätzung beruht auf der *generalisierte Momentenmethode* oder auf der *empirical likelihood*-Methode; für technische Details sei auf den Artikel verwiesen. Mit ihrem Vorschlag, explizit eine Balance der Kovariablen in der Schätzung mit berücksichtigen zu können, reihen sich die Autoren dem Vorschlag von Hainmueller (2011) ein,

der einen propensity score-Schätzer vorschlägt, der die Entropie der Verteilungen balanciert.

1.11 Propensity Score-Methoden

Einleitung

Es liegen verschiedene Adjustierungsmethoden vor, die den propensity score, $e(\vec{x}_i)$, nutzen, um rückwirkend in einer observational study diejenige Balance in den Kovariablen zu rekonstruieren, die sich bei Randomisierung ergeben hätte und somit eine unkonfundierte Schätzung des ACE zu ermöglichen. Der basale Ansatz ist allen folgend aufgeführten Methoden immanent: Konditioniert an $e(\vec{x}_i)$ sind die Verteilungen der Kovariablen homogen verteilt und das treatment-assignment gilt als unabhängig von den response-Variablen, d.h. das treatment-assignment ist *strongly ignorable*; damit folgt das Design konditional an $e(\vec{x}_i)$ einem „mini-randomized experiment“ (Stuart, 2010, S.12). Somit lässt sich bedingt an $e(\vec{x}_i)$ der ACE identifizieren und schätzen, wobei für eine erwartungstreue und konsistente Schätzung des ACE unter Verwendung der folgenden Adjustierungsmethoden eine korrekte Spezifikation des propensity score-Modells vorausgesetzt wird, da in praxi $e(\vec{x}_i)$ ein unbekannter Skalar ist. Die Relevanz einer korrekten Spezifikation des propensity score-Modell wurde mehrfach in Simulationsstudien empirisch aufgewiesen (z.B., Dehejia & Wahba, 2002; Drake, 1993; Zhao, 2004): gegeben einer Fehlspezifikation des propensity score-Modells resultieren verzerrte, adjustierte Schätzungen des ACE. Es zeigt sich allgemein, dass die Hinzunahme von Variablen in das propensity score-Modell, die keinen oder nur einen geringfügigen Effekt auf das treatment-assignment besitzen, keinerlei Auswirkungen auf die Schätzung des ACE hat, jedoch die Exklusion von Variablen, die entscheidend das treatment-assignment determinieren (sog. *omitted variables*-Problem), den vorliegenden bias, der auf die Konfundierung der Kovariablen zurückzuführen ist, erhöht. Folgend vorzustellende Adjustierungsmethoden setzen eine korrekte Spezifikation des propensity scores voraus, damit in Folge dessen der ACE erwartungstreu und konsistent geschätzt werden kann.

1.11.1 Matching

Einführung: Exaktes Matching

In einer Reihe von Veröffentlichungen (z.B., Rosenbaum, 2010; Rosenbaum & Rubin, 1983; Rubin, 1977) wird auf die Motivation des matchings durch das Modell des *ideal matchings* bzw. *exakten matchings* hingewiesen. Anhand eines two-trials Beispiel soll der Kern aufgewiesen werden; es werden folgende Annahmen getroffen:

a) Es liegt ein strongly ignorable treatment-assignment vor, so dass

$$\begin{aligned}
&\Pr(S_1 = s_1, \ldots, S_N = s_N \mid y_{t,1}, y_{c,1}, \vec{x}_1, \vec{z}_1, \ldots, y_{t,N}, y_{c,N}, \vec{x}_N, \vec{z}_N) \\
=&\Pr(S_1 = s_1, \ldots, S_N = s_N \mid \vec{x}_1, \ldots, \vec{x}_N) \\
=&\prod_{i=1}^{N} \pi_i^{s_i}(1-\pi_i)^{(1-s_i)} \qquad \pi_i = \Pr(S_i = 1 \mid \vec{x}_i),\ 0 < \pi_i < 1.
\end{aligned}$$

b) Es lassen sich zwei Einheiten i und i' finden, für die

b.1) $\vec{x}_i = \vec{x}_{i'}$

b.2) $s_i + s_{i'} = 1$

b.3) $\pi_i = \pi_{i'}$

Es folgt für i und i' damit:

$$\begin{aligned}
&\Pr(S_i = s_i, S_{i'} = s_{i'} \mid \vec{x}_i, \vec{x}_{i'}) \\
=&\pi_i^{s_i}(1-\pi_i)^{(1-s_i)}\pi_{i'}^{s_{i'}}(1-\pi_{i'})^{(1-s_{i'})} \\
=&\pi_i^{s_i+s_{i'}}(1-\pi_i)^{(1-s_i)+(1-s_{i'})}
\end{aligned}$$

Für die *Chance*, dass i das treatment erhält und i' in die Kontrollbedingung selegiert wird, folgt:

$$\begin{aligned}
&\gamma(S_i = 1, S_{i'} = 0 \mid \vec{x}_i, \vec{x}_{i'}) \\
=&\frac{\Pr(S_i = 1, S_{i'} = 0 \mid \vec{x}_i, \vec{x}_{i'})}{\Pr(s_i + s_{i'} = 1 \mid \vec{x}_i, \vec{x}_{i'})} \\
=&\frac{\pi_i \cdot (1-\pi_i)}{\pi_i \cdot (1-\pi_i) + \pi_i \cdot (1-\pi_i)} \\
=&\frac{\pi_i \cdot (1-\pi_i)}{2 \cdot (\pi_i \cdot (1-\pi_i))} \\
=&\frac{1}{2}
\end{aligned}$$

Demnach lässt sich für i und i' mit $\vec{x}_i = \vec{x}_{i'}$ - sie gelten folgend in ihren Kovariablenvektoren als *matched* - aufweisen, dass das treatment-assignment gegeben $\vec{x}_i$ randomisiert erfolgt: i und i' besitzen bedingt an $\vec{x}_i$ die selbe Chance, das treatment zu erhalten.

Grundlegend wird beim exakten matching jeder der N_t-Einheiten des treatments als match eine Einheit i' aus c, d.h. $i' \in \{i \mid i \in U \cap (S_i = 0)\}$, zugeordnet, die dieser Einheit in ihrem Kovariablenvektor bzw. propensity score gleicht. Entsprechend aller vorangestellten Überlegungen folgt dem exakten matching damit:

$$F(\mathbf{X} \mid S_i = 1) = F(\mathbf{X} \mid S_i = 0, M_i = 1) \text{ sowie}$$
$$E(Y_t \mid S_i = 1) = E(Y_t \mid S_i = 0, M_i = 1),$$

wobei folgend M_i als matching-Indikator definiert ist mit:

$$M_i = \begin{cases} 1, & \text{wenn } i, i : S_i = 0, \text{ gematched} \\ 0, & \text{sonst.} \end{cases}$$

Ausgehend von dieser Motivation sollen damit folgend die in praxi relevanten matching-Verfahren und Schätzstatistiken für den treatment-Effekt dargestellt werden.

Matching: Allgemeine Definition und Ablauf

Als *matching* werden im Allgemeinen statistische Methoden zusammengefasst, deren Anwendungen auf eine identische Verteilung der Kovariablen in den Experimentalbedingungen abzielen und dem folgend ein strongly ignorable treatment-assignment plausibel machen (Stuart, 2010). Bei einem *matched pair* handelt es sich demnach um zwei Einheiten, i und i', wobei i' zur Teilmenge der N_c-Einheiten der Kontrollbedingung, i entsprechend zur Teilmenge der N_t-Einheiten des treatments gehört, das nach Anwendung eines matching-Algorithmus als identisch, oder ähnlich, in ihren Kovariablenvektoren, $\vec{x}_i$ und $\vec{x}_{i'}$, klassifiziert wird (z.B., Imbens & Rubin, 2012; Rosenbaum & Rubin, 1985a, 1985b). Es besteht beim matching grundlegend die Möglichkeit, dass exakt ein i' einem einzigen i als match zugeordnet wird, sog. $1:1-$matching, oder mehrere i' einem einzelnen i, sog. $1:k-$matching; als Resultat liegen stets N_t-Paare oder Gruppen von matched Einheiten vor, wobei alle matched i' eine neue Kontrollgruppe definieren, die den Einheiten aus t in ihren Kovariablenvektoren gleichen. Dementsprechend werden als propensity score-matching diejenigen Methoden zusammengefasst, die statistische

Einheiten aus c, die identische oder ähnliche $e(\vec{x}_i)$ wie i im treatment besitzen, zu einzelnen Teilstichproben poolen.

Eine provisorische Einteilung der matching-Methoden wird vorgeschlagen (Guo & Fraser, 2010), wobei eine Vielzahl von Abwandlungen und Kombinationen der Methoden diskutiert wird oder vollständig abweichende Methoden entworfen werden (z.B., Abadie & Imbens, 2002, 2011; Heckman et al., 1997; Imai et al., 2008).

1. greedy matching:
   (a) exaktes matching
   (b) Mahalanobis-Distanzen matching
   (c) Mahalanobis-Distanzen matching mit $e(\vec{x}_i)$
   (d) caliper-matching
   (e) nearest neighbor matching
   (f) nearest neighbor matching innerhalb eines calipers
   (g) nearest available Mahalanobis-Distanz innerhalb eines calipers durch $e(\vec{x}_i)$
2. optimal matching
3. fine balance-Methoden

Das gesamte matching-Verfahren beinhaltet mehrere Schritte, die in einer Schätzung des ACE oder des ATT münden:

1. Auswahl der Kovariablen, mittels derer die Einheiten gematched werden bzw. Auswahl der Kovariablen zur Modellierung des propensity scores
2. Messung der Distanzen der Einheiten
3. Anwendung des matching-Algorithmus
4. Schätzung des treatment-Effekts

Folgend sollen die gebräuchlichsten Distanzmaße und die verbundenen matching-Algorithmen vorgestellt werden; für eine vollständige Übersicht sei an dieser Stelle auf Literatur verwiesen,

die reviews über dieses Themenfeld darstellen (z.B., Guo & Fraser, 2010; Imbens, 2004; Rosenbaum, 1995, 2010; Stuart, 2010). Der Schwerpunkt der Aufführungen liegt der Kürze halber beim *greedy matching*, das in praxi die meiste Anwendung findet (z.B., Dehejia & Wahba, 2002; Stuart, 2010).

Distanzmaße

Gematched mit i gilt eine Einheit $i^{'}$, wenn diese, der Kontrollbedingung zugeordnet, identische oder vergleichbare $\vec{x}_i$ bzw. $e(\vec{x}_i)$ wie i besitzt; damit gilt es zunächst, um zwei Einheiten zu matchen, ein Maß zu definieren, welches die Ähnlichkeit zweier Einheiten indiziert, folgend *Distanzmaß*. Es finden sich in der Literatur verschiedene Distanzmaße (z.B., Guo & Fraser, 2010; Imbens, 2004; Rosenbaum & Rubin, 1985b; Rosenbaum, 2010; Stuart, 2010), die folgend kurz dargestellt werden sollen; formal wird an dieser Stelle das Distanzmaß $D_{i,i'}$ eingeführt.

Das vorangestellte *exakte matching* setzte voraus, dass an i und i' identische $\vec{x}_i$ bzw. $e(\vec{x}_i)$ beobachtet werden - es handelt sich damit um das restriktivste Distanzmaß mit

$$D_{i,i'} = \begin{cases} 0, & \text{wenn } \vec{x}_i = \vec{x}_{i'} \\ \infty & \text{wenn } \vec{x}_i \neq \vec{x}_{i'}. \end{cases} \tag{25}$$

Ein exaktes matching gegeben $e(\vec{x}_i)$ würde als Distanzmaß

$$D_{i,i'} = \begin{cases} 0, & \text{wenn } e(\vec{x}_i) = e(\vec{x}_{i'}) \\ \infty & \text{wenn } e(\vec{x}_i) \neq e(\vec{x}_{i'}) \end{cases} \tag{26}$$

entsprechend nutzen.

Dieses restriktive Vorgehen führt in seiner Anwendung zu einem grundlegenden Problem, das dem Phänomen der *curse of dimensionality* zugeordnet werden kann: mit zunehmender Anzahl an Kovariablen konvergiert die Wahrscheinlichkeit für einen exakten match gegen null. Im einfachsten Falle liegen bei p binären Kovariablen 2^p-Kovariablenvektoren $\vec{x}_i$ vor und ein exaktes matching wird damit explizit bei kleinen Kontrollgruppen und großem p unwahrscheinlich. Diese Problematik lässt sich unmittelbar auf stetige Kovariablen ausweiten: Abadie und Imbens (2002) zeigen in ihrem ersten Lemma, dass die Wahrscheinlichkeit für ein exaktes matching mit einer (oder mehreren) stetigen Kovariable(n) gegen null konvergiert, so dass ein exaktes propensity score-matching, wenn vollständig stetig gemessen, in kleinen Kontrollgruppen ebenfalls schwer realisierbar wird. In Konsequenz lässt sich für einen Anteil an Einheiten in

t kein entsprechendes match aus c finden, so dass das matching *unvollständig* bzw. *incomplete* ist und weiterhin entsprechende Konfundierungen durch die Kovariablen resultieren können (Rosenbaum & Rubin, 1985a):

Dem unvollständigen matching resultiert stets ein bias, formal

$$E(\mathbf{X} \mid S_i = 1) - E(\mathbf{X} \mid S_i = 0, M_i = 1) \neq 0,$$

der einem unvollständigen, exakten matching folgend größer ist, als wenn *approximativ* gematched wird, demnach ähnliche und nicht identische $\vec{x}_i$ bzw. $e(\vec{x}_i)$ gematched werden (Rosenbaum & Rubin, 1985a). Ein Beispiel für ein approximatives matching ist die sog. *Stratifikation*, die im nächsten Kapitel definiert wird.

Als weiteres Distanzmaß lässt sich die *Euklidische Distanz* definieren (Imbens, 2004):

$$D_{i,i'} = \sqrt{(\vec{x}_i - \vec{x}_{i'})^t(\vec{x}_i - \vec{x}_{i'})}. \tag{27}$$

Es ist bei der Ermittlung der Euklidischen Distanz ersichtlich, dass jede Koordinate in $\vec{x}_i$ und $\vec{x}_{i'}$ gleichwertig gewogen wird. Damit ist ein matching mit dem Euklidischen Distanzmaß stets abhängig von der Metrik der erhobenen Kovariablen, da Kovariablen mit großen Standardabweichungen die ermittelten Distanzen dominieren. Zusätzlich lässt sich zeigen, dass mögliche Korrelationen zwischen den Variablen die Distanzen verzerren. Als Lösung dieser Probleme lässt sich basierend auf der Euklidischen Distanz die *Mahalanobis*-Distanz definieren, die aus einer Standardisierung der Kovariablen hervorgeht, damit skaleninvariant ist und weiterhin die Kovarianzen zwischen den Kovariablen berücksichtigt (z.B., Baser, 2006; Guo & Fraser, 2010; Imbens, 2004; Rubin, 1976):

$$D_{i,i'} = \sqrt{(\vec{x}_i - \vec{x}_{i'})^t \mathbf{\Sigma}^{-1} (\vec{x}_i - \vec{x}_{i'})}, \tag{28}$$

wobei $\mathbf{\Sigma}^{-1}$: invertierte Varianz-Kovarianzmatrix der Kovariablen.

Für das matching mit der Mahalanobis-Distanz werden vor Bestimmung der Distanzen die statistischen Einheiten dem Zufall nach angeordnet, anschließend werden die Distanzen für die erste Einheit aus t zu allen Einheiten in c bestimmt (z.B., Baser, 2006; Guo & Fraser, 2010; Rosenbaum & Rubin, 1985b). Zumeist wird der Bestimmung der Distanzen folgend ein *nearest neighbor matching* durchgeführt: ein match ist damit definiert als diejenige Einheit i', die zu i die geringste Mahalanobis-Distanz aufweist. Beide Einheiten bilden damit eine *neighborhood*, $C(d_{i,i'})$, die durch das Paar i und i' definiert wird, das zueinander die geringste

absolute Differenz in ihren Mahalanobis-Distanzen aufweist - der matching-Algorithmus für ein nearest neighbor matching lautet entsprechend:

$$C(d_{i,i'}) := \min_{i'} \mid d_{i,i'} \mid \tag{29}$$

Wird ein treated i genau mit einem i' aus der Kontrollbedingung gematched, handelt es sich um ein sog. 1:1-matching; befinden sich in der neighborhood $C(d_{i,i'})$ $k-$Einheiten aus c, handelt es sich um ein 1:k-matching. Typischerweise werden nach dem match beide Einheiten bzw. die $(1+k)-$Einheiten aus dem pool der potentiellen matches entfernt und der matching-Algorithmus wird für alle übrig gebliebenen Einheiten repetiert (*matching ohne Zurücklegen*; z.B., Abadie und Imbens (2002); Rosenbaum (1995, 2010)). Abadie und Imbens (2002) betonen, dass ein matching mit Zurücklegen einhergeht mit genaueren matches, da i' mehrfach als match fungieren kann, insofern es zu mehreren treated Einheiten dieselbe Ähnlichkeit aufweist; in Konsequenz resultiert eine größere bias-Reduktion bei der Schätzung des ACE, jedoch verbunden mit einer größerem Standardfehler. Rubin (1976) führt aus, dass ein matching mit der Mahalanobis-Metrik bei multivariater Normalverteilung der Variablen in $\mathbf{X}$ mit identischen Varianz-Kovarianzmatrizen, $\mathbf{\Sigma}_t$ und $\mathbf{\Sigma}_c$, einhergeht mit einer *equal-percent bias reduction* (EPBR), d.h. einer identischen prozentualen bias-Reduktion für alle Koordinaten in $\vec{x}_i$:

Der initiale bias definiert ist als $E(\mathbf{X} \mid S_i = 1) - E(\mathbf{X} \mid S_i = 0)$, der bestehende bias nach dem matching-Prozedere als $E(\mathbf{X} \mid S_i = 1) - E(\mathbf{X} \mid S_i = 0, M_i = 1)$.

Formell ist damit eine EPBR definiert als:

$$E(\mathbf{X} \mid S_i = 1) - E(\mathbf{X} \mid S_i = 0, M_i = 1) = \lambda \left(E(\mathbf{X} \mid S_i = 1) - E(\mathbf{X} \mid S_i = 0) \right), \quad 0 \leq \lambda \leq 1.$$

Es liegen Vorschläge vor, zusätzlich zu den Kovariablen, für die gematched werden soll, $e(\vec{x}_i)$ mit einzubeziehen bei der Bestimmung der Mahalanobis-Distanzen, da dieser per definitionem als balancing score sämtliche Kovariablen des propensity score-Modells balanciert (z.B., Baser, 2006; Rosenbaum & Rubin, 1985b; Stuart, 2010). Entsprechend kann die Dimensionalität des Kovariablenvektors durch eine vorherige Modellierung des propensity scores reduziert werden und damit mögliche Probleme, die im Zusammenhang mit dem *curse of dimensionality* stehen, umgangen werden. Das Distanzmaß lautet damit:

$$D_{i,i'} = \sqrt{(\vec{u}_i - \vec{u}_{i'})^t \mathbf{\Sigma}^{-1} (\vec{u}_i - \vec{u}_{i'})}, \tag{30}$$

wobei $\vec{u}_i^t = \{\vec{x}_i, e(\vec{x}_i)\}$ oder alternativ $\vec{u}_i^t = \{\vec{x}_i, \text{logit}(e(\vec{x}_i))\}$. $\mathbf{\Sigma}^{-1}$ ist weiterhin definiert wie in (28). Empirisch zeigt sich, dass ein nearest neighbor matching mit dem in (30) definierten Ma-

halanobis-Distanzmaß in Hinblick auf die Kovariablenbalance dem einfachen nearest neighbor matching mittels $e(\vec{x}_i)$ überlegen ist, jedoch die propensity scores nach dem matching mit der Mahalanobis-Distanz größere Differenzen zwischen den Gruppen aufweisen als gegeben dem nearest neighbor matching mit $e(\vec{x}_i)$ (Rosenbaum & Rubin, 1985b). In Anlehnung an die equal-percent bias reduction wird bei Verwendung des propensity scores zur Bestimmung der Mahalanobis-Distanzen, vgl. (30), angeraten, die zugehörigen logits zu nutzen (z.B., Guo & Fraser, 2010; Rosenbaum & Rubin, 1985b).

Für ein mögliches propensity score - matching werden weitere Distanzmaße und Methoden diskutiert:

$$D_{i,i'} = \mid e(\vec{x}_i) - e(\vec{x}_{i'}) \mid \tag{31}$$

indiziert die absolute Differenz zwischen den Einheiten i und i' in ihren propensity scores. Im restriktivesten Falle, analog zu (25), gelten zwei Einheiten in ihrem propensity score gematched, wenn $\mid e(\vec{x}_i) - e(\vec{x}_{i'}) \mid = 0$. Das *nearest neighbor matching* nutzt die in (31) definierten Abstände zwischen zwei Einheiten und matched diese, wenn die gemessene absolute Distanz zwischen den propensity scores minimal ist.

Wie oben angedeutet, folgt einer Linearisierung des propensity scores, beispielsweise durch Ermittlung der zugehörigen logits, im Verhältnis zu (31) eine größere prozentuale bias-Reduktion der Kovariablen (Rosenbaum & Rubin, 1985b); vorgeschlagen als Distanzmaß wird damit:

$$D_{i,i'} = \mid \text{logit}\,(e(\vec{x}_i)) - \text{logit}\,(e(\vec{x}_{i'})) \mid \tag{32}$$

Eine weitere Möglichkeit, die Ähnlichkeit zweier Einheiten zu messen, ergibt sich durch das *caliper-matching*, wobei alle bisherigen Distanzmaße vereinbar sind mit dieser Methodik. Das caliper-matching wurde ausgehend von der Problematik entwickelt, dass ein nearest neighbor match durch die minimale Distanz zweier Einheiten definiert ist, jedoch keine Restriktion vorliegt, wie groß die Distanz maximal sein darf. Als caliper ϵ wird ein, vor dem Messen der Distanzen festgelegtes Toleranzmaß bezeichnet, so dass nur Distanzen berücksichtigt werden, für die, z.B. im Falle der Absolutdifferenzen für $e(\vec{x}_i)$, gilt:

$$D_{i,i'} = \begin{cases} \mid e(\vec{x}_i) - e(\vec{x}_{i'}) \mid & \text{wenn } \mid e(\vec{x}_i) - e(\vec{x}_{i'}) \mid \leq \epsilon \\ \infty & \text{wenn } \mid e(\vec{x}_i) - e(\vec{x}_{i'}) \mid > \epsilon \end{cases} \tag{33}$$

Rosenbaum und Rubin (1985b) sowie Rubin und Thomas (2000) schlagen als weiteres Distanzmaß das *nearest available Mahalanobis metric matchin withing calipers defined by the propensity*

score, formal:

$$D_{i,i'} = \begin{cases} \sqrt{(\vec{u}_i - \vec{u}_{i'})^t \boldsymbol{\Sigma}^{-1} (\vec{u}_i - \vec{u}_{i'})} & \text{wenn } \mid \text{logit}(e(\vec{x}_i)) - \text{logit}(e(\vec{x}_{i'})) \mid \leq \epsilon \\ \infty & \text{wenn } \mid \text{logit}(e(\vec{x}_i)) - \text{logit}(e(\vec{x}_{i'})) \mid > \epsilon, \end{cases} \tag{34}$$

wobei $\vec{u}_i^t = \{\vec{x}_i, \text{logit}(e(\vec{x}_i))\}$. Damit werden im ersten Schritt diejenigen Einheiten aus c als potentielle matches gesucht, deren absoluten Abstände der propensity scores oder logits innerhalb des festgelegten calipers liegen, im Folgeschritt wird dasjenige i' innerhalb des calipers als match bestimmt, dass eine minimale Mahalanobis-Distanz (34) aufweist. Empirisch zeigt sich, dass ein matching-Algorithmus nach (34) sowohl die Kovariablen, die bei der Bestimmung der Mahalanobis-Distanzen eine Rolle spielen, als auch die $e(\vec{x}_i)$ balanciert und damit der Methodik des *nearest neighbor matching* gegeben $e(\vec{x}_i)$ als auch dem *nearest neighbor matching mittels der Mahalanobis-Distanzen inklusive* $e(\vec{x}_i)$ überlegen ist (Rosenbaum & Rubin, 1985b). In Anlehnung an Cochran und Rubin (1973) zeigt sich, dass, wenn die Varianz der logits in t doppelt so groß ist wie in der Kontrollbedingung, d.h. $\sigma_t^2/\sigma_c^2 = 2$, ein caliper mit $\epsilon = 0.2\sigma$ - wobei $\sigma = \sqrt{\sigma_t^2 + \sigma_c^2}$ - ca. 98% der Verzerrungen in den Kovariablen balanciert.

Schätzung des Treatment-Effekts

Dem matching-Algorithmus folgt, zumeist nach einer Diagnose der resultierenden Kovariablenbalance in der neuen Stichprobe der matches, die Schätzung des treatment-Effektes (z.B., Guo & Fraser, 2010; Imai et al., 2008; Rosenbaum, 2010; Stuart, 2010). Eine mögliche Überprüfung der Kovariablenbalance erfolgt beispielsweise durch Boxplots der Kovariablen, abgetragen getrennt für die treatment- und Kontrollgruppe, in der Stichprobe derer, die gematched worden sind.

Ausgehend von einem *exakten matching* lässt sich, irrelevant, ob mittels $\vec{x}_i$ oder $e(\vec{x}_i)$ gematched wird, der ATT (s. (9)) als treatment-Effekt identifizieren (z.B., Imbens & Rubin, 2012; Rosenbaum & Rubin, 1985a). Es lässt sich N_t-resultierenden Paare entsprechend definieren:

$$\text{ATT} = \frac{1}{N_t} \left(\sum_{i:S_i=1} y_{i,t} - \sum_{i:i \in \{S_i=0, M_i=1\}} y_{i,c} \right) \tag{35}$$

Als Schätzstatistik lässt sich entsprechend die Differenz der Mittelwerte beider Experimental-

gruppen in der matched Stichprobe, d.h. $\bar{Y}_t - \bar{Y}_{c|M_i=1}$, definieren:

$$\hat{\text{ATT}} = \bar{Y}_t - \bar{Y}_{c|M_i=1} = \frac{1}{n_t} \sum_{i:i\in\{I_i=1,S_i=1\}} y_{i,t} - \frac{1}{n_t} \sum_{i:i\in\{I_i=1,S_i=0,M_i=1\}} y_{i,c} \tag{36}$$

$$\hat{\sigma}^2_{\text{ATT}} = \frac{\hat{\sigma}^2_t}{n_t} + \frac{\hat{\sigma}^2_{c|M_i=1}}{n_t}. \tag{37}$$

Dem matching mit der *Mahalanobis*-Distanz folgend lassen sich unterschiedliche treatment-Effekte identifizieren: Insofern der ACE identifiziert werden soll, ist $\mathbf{\Sigma}^{-1}$ in (28) festgelegt durch die Varianzen und Kovarianzen der Kontrollgruppe c (z.B., Guo & Fraser, 2010; Rosenbaum & Rubin, 1985b; Stuart, 2010). Stuart (2010) zeigt, dass zur Schätzung des ATT die gepoolten Varianzen und Kovarianzen für die matched Einheiten aus t sowie der gesamten Kontrollgruppe benötigt werden.

In der Literatur werden eine Vielzahl von weiteren matching-Algorithmen diskutiert, die in dieser Arbeit auf Grund der Länge nicht weiter vorgestellt werden können; es sei an dieser Stelle auf einige Übersichtsarbeiten verwiesen (z.B., Hansen, 2004; Heckman et al., 1997; Guo & Fraser, 2010; Rosenbaum, 1995, 2010; Stuart, 2010).

1.11.2 Propensity Score-Stratifikation

Die Methodik der propensity score-Stratifikation lässt sich als Spezialfall des *approximativen matchings* ansehen; ein stratum entspricht der Zusammenfassung von statistischen Einheiten, die mit ähnlichen propensity scores gematched werden.

Es werden bei der Stratifikation $v-$strata, $j = 1, \ldots, v$, gebildet, wobei die strata benachbarte, nach unten offene, nach oben geschlossene Intervalle darstellen; in praxi werden die Obergrenzen der strata durch die Quantile der propensity scores bestimmt (Rosenbaum & Rubin, 1984). An erster Stelle wird für eine Stratifikation die Ordnungsstatistik der propensity scores bestimmt, so dass

$$e(\vec{x}_1) \leq \ldots \leq e(\vec{x}_i) \leq \ldots \leq e(\vec{x}_N).$$

Für eine Quantil-Stratifikation werden als Obergrenzen der strata diejenigen $e(\vec{x}_i)$ bestimmt, die mit den $p-$Quantilen zusammenfallen, so dass das $j-$te stratum dem Intervall $Q_j := (q_{j-1}; q_j]$ entspricht, wobei q_j übereinstimmt mit demjenigen $e(\vec{x}_i)$, der den Anteil p, $p = \frac{j}{v}$, an statistischen Einheiten mit $e(\vec{x}_i) \leq q_j$ einschließt. Definitionsgemäß gilt $q_0 = 0$ sowie $q_v = 1$.

Innerhalb eines $j-$ten stratums lässt sich der stratumsspezifische ACE durch die Statistik

$$\hat{\text{ACE}}_j := \bar{Y}_{t,j} - \bar{Y}_{c,j} = \left(\frac{1}{n_{t,j}} \sum_{i \in \{I_i=1, S_i=1, I(e(\vec{x}_i) \in Q_j)\}} y_{i,t} \right) - \left(\frac{1}{n_{c,j}} \sum_{i \in \{I_i=1, S_i=0, I(e(\vec{x}_i) \in Q_j)\}} y_{i,c} \right) \quad (38)$$

schätzen. Der unkonditionierte ACE lässt sich entsprechend durch das gewogene Mittel der stratumsspezifischen Punktschätzungen schätzen, so dass:

$$\hat{\text{ACE}} := \sum_{j=1}^{v} \left(\frac{n_j}{n} \right) \cdot (\bar{Y}_{t,j} - \bar{Y}_{c,j}), \quad (39)$$

wobei n_j : die Anzahl an statistischen Einheiten im $j-$ten stratum und $\frac{n_j}{n}$: Anteil der gesampleden Einheiten, die dem $j-$stratum zugehörig sind. Gegeben einer Quantil-Stratifikation lässt sich $\frac{n_j}{n}$ ersetzen durch $\frac{1}{v}$. Die geschätzte Varianz der Schätzstatistik im $j-$ten stratum entspricht

$$\hat{\sigma}^2_{\hat{\tau}_j} := \hat{\sigma}^2_{\bar{Y}_{t,j} - \bar{Y}_{c,j}}, \quad (40)$$

eine stratumsunabhängige Varianzschätzung der Statistik wird entsprechend über alle v strata ermöglicht durch

$$\hat{\sigma}^2_{\hat{\tau}} := \sum_{j=1}^{v} \left(\frac{n_j}{n} \right)^2 \cdot \hat{\sigma}^2_{\bar{Y}_{t,j} - \bar{Y}_{c,j}}. \quad (41)$$

In der Literatur findet sich zumeist der Vorschlag, dass in praxi eine *Quintil-Stratifikation* vorgenommen werden sollte, um den Effekt der konfundierenden Kovariablen zu minimieren (z.B., Rosenbaum & Rubin, 1983; Rosenbaum, 1984a; Stuart, 2010); in Anlehnung an Cochran (1968) lässt sich zeigen, dass eine Quintil-Stratifikation gegeben der propensity scores zu einer Reduktion von ca. 90% des bias, der auf die konfundierenden Kovariablen zurückzuführen ist, bei der Schätzung des ACE führt (z.B., Rosenbaum & Rubin, 1984, 1985b).

Lediglich in einem Range von $0 < e(\vec{x}_i) < 1$ ist eine Schätzung des ACE durch die Gruppenmittelwerte möglich; in den Extremstrata Q_1 und Q_v können entsprechend statistische Einheiten zugehörig sein, die allesamt einer der beiden Experimentalbedingungen zugewiesen sind. Lassen sich entsprechende strata beobachten, in denen die Einheiten eindeutig einem $k \in K$ zugewiesen sind, werden diese von der Schätzung des ACE ausgeschlossen, da eine Differenz der Gruppenmittelwerte nicht definiert ist (D' Agostino, 1998).

Lunceford und Davidian (2004) führen in ihrem Artikel formal die *large sample*-Eigenschaften einer propensity score-Stratifikation auf, auf die folgend kurz hingewiesen werden sollen; für Formalitäten und Beweise sei auf den Artikel hingewiesen. Die Autoren führen auf, dass der in (39)

definierte Schätzer stets eine Verzerrung bei der Schätzung des ACE aufweist und inkonsistent ist, selbst dann, wenn propensity score-Modells korrekt spezifiziert wurde. Diese Verzerrung lässt sich unmittelbar begründen, da bei einer Stratifikation gleichmäßig Anteile von statistischen Einheiten mit ähnlichen, jedoch nicht notwendigerweise identischen, propensity scores zusammengefasst werden. Dem einhergehend lässt sich innerhalb eines stratums eine Varianz an propensity scores festhalten, die einhergeht mit einer weiterhin bestehenden Konfundierung der Kovariablen (sog. *residual within-stratum confounding*).

1.11.3 Propensity Score-Weighting

Idee des Horvitz-Thompson-Schätzers

Unter der Voraussetzung, dass alle $N-$Einheiten in U die selbe Auswahlwahrscheinlichkeit besitzen, so dass: $\Pr(I_i = 1) = \frac{n}{N}$, $\forall i \in U$, zeigt sich, dass die Schätzstatistik $g(Y_1 \ldots, Y_n) = \bar{Y}$ erwartungstreu, konsistent und effizient den Erwartungswert $E(Y_i) = E(y_i \cdot I_i)$ schätzt; damit: $E(\bar{Y}) = E(Y_i)$. Es lässt sich aufweisen, dass gegeben unterschiedlicher Auswahlwahrscheinlichkeiten, d.h. $\Pr(I_i = 1) \neq \Pr(I_{i'} = 1) \neq \frac{n}{N}$, die Schätzfunktion $\bar{Y}$ einen bias bei der Schätzung von $E(Y_i)$ aufweist; durch eine nachträgliche Gewichtung der Realisierung $Y_i = y_i$ kann diese Verzerrung ausgeglichen werden, da bei der Ermittlung des Punktschätzers entsprechend die unterschiedlichen Auswahlwahrscheinlichkeiten berücksichtigt würden. Horvitz und Thompson (1952) zeigten, dass eine Gewichtung mittels der invertierten Auswahlwahrscheinlichkeit, $\frac{1}{\Pr(I_i=1)}$, erfolgen kann und definierten einen erwartungstreuen Punktschätzer für $E(Y_i)$, der gültig für eine breite Klasse von möglichen Stichprobendesigns, deren Gemeinsamkeit eine Stichprobenziehung aus U ohne Zurücklegen und ungleichen Auswahlwahrscheinlichkeiten ist, ein. Dieser Punktschätzer ist definiert als *Horvitz-Thompson-Schätzer* (Kauermann & Küchenhoff, 2011); formal gilt:

$$\bar{Y}_{HT} = \frac{1}{N} \sum_{i:I_i=1} \frac{y_i}{\Pr(I_i = 1)} \tag{42}$$

Die Erwartungstreue dieses Schätzers für $E(Y_i)$ lässt sich einfach zeigen:

$$E(\bar{Y}_{HT}) = E\left(\frac{1}{N}\sum_{i=1}^{N} \frac{y_i}{\Pr(I_i = 1)} \cdot I_i\right) = \frac{1}{N}\sum_{i=1}^{N} \frac{y_i}{\Pr(I_i = 1)} \cdot E(I_i)$$

$$= \frac{1}{N}\sum_{i=1}^{N} \frac{y_i}{\Pr(I_i = 1)} \cdot \Pr(I_i = 1) = \frac{1}{N}\sum_{i=1}^{N} y_i = E(Y_i)$$

Propensity Score-Weighting

Robins, Hernán und Brumback (2000) implementierten die Idee des *inverse-probability-weighting*, das dem Horvitz-Thompson-Schätzer zu Grunde liegt, in den Kontext des *missing data*-Problems und definierten in diesem Zusammenhang mehrere konsistente und semiparametrische Schätzer, die in ihrem Schätzprinzip problemlos in den Kontext der kausalen Inferenz übertragen werden können: wie zu zeigen ist, lässt sich ein erwartungstreuer Schätzer für den unbekannten Erwartungswert $E(Y_t)$ durch eine Gewichtung mittels des invertierten propensity scores konstruieren; vice versa für $E(Y_c)$. Für die Herleitung werden folgende Annahmen benötigt:

a) $S_i(1 - S_i) = 0$

b) $E\left(\frac{S_i \cdot Y_i}{e(\vec{x}_i)}\right) \overset{\text{SUTVA}}{=} E\left(\frac{S_i \cdot y_{i,t}}{e(\vec{x}_i)}\right)$

Es folgt unter der Anwendung des *law of iterated expectations*:

$$E\left(\frac{S_i \cdot Y_i}{e(\vec{x}_i)}\right) = E\left[E\left(\frac{I(S_i = 1) \cdot y_{i,t}}{e(\vec{x}_i)}\right)\middle| y_{i,t}, \vec{x}_i\right] = E\left(\frac{y_{i,t}}{e(\vec{x}_i)} E(I(S_i = 1)\middle| y_{i,t}, \vec{x}_i)\right) = E(Y_t),$$

wobei: $E(I(S_i = 1) \mid y_{i,t}, \vec{x}_i) = \Pr(S_i = 1 \mid y_{i,t}, \vec{x}_i) = \Pr(S_i = 1 \mid \vec{x}_i) = e(\vec{x}_i)$ und $S_i = I(S_i = 1)$. Unter Anwendung selbiger Regeln folgt für die Kontrollbedingung:

$$E\left(\frac{(1 - S_i) \cdot Y_i}{(1 - e(\vec{x}_i))}\right) = E(Y_c)$$

Entsprechend dieser Herleitung lässt sich unmittelbar als Schätzstatistik für den ACE der IPW_1-Schätzer konstruieren

$$IPW_1 := \frac{1}{n} \sum_{i:I_i=1} \frac{S_i \cdot Y_i}{e(\vec{x}_i)} - \frac{1}{n} \sum_{i:I_i=1} \frac{(1 - S_i) \cdot Y_i}{(1 - e(\vec{x}_i))}, \tag{43}$$

der in einer Vielzahl an Veröffentlichungen Verwendung findet (z.B., Austin, 2011; Lunceford & Davidian, 2004; Rosenbaum, 1987; Stuart, 2010).

Durch die zusätzlichen Annahmen

c) $E\left(\frac{S_i}{e(\vec{x}_i)}\right) = E\left(\frac{E(S_i|\vec{x}_i)}{e(\vec{x}_i)}\right) = 1$

d) $E\left(\frac{(1-S_i)}{(1-e(\vec{x}_i))}\right) = 1$

lässt sich des weiteren der IPW_2-Schätzer definieren:

$$IPW_2 := \frac{1}{\left(\sum_{i:I_i=1} \frac{S_i}{e(\vec{x}_i)}\right)} \sum_{i:I_i=1} \frac{S_i \cdot Y_i}{e(\vec{x}_i)} - \frac{1}{\left(\frac{1-S_i}{1-e(\vec{x}_i)}\right)} \sum_{i:I_i=1} \frac{(1 - S_i) \cdot Y_i}{1 - e(\vec{x}_i)}. \tag{44}$$

Der Zusammenhang zur *missing data*-Problematik lässt sich leicht herstellen: Ausgangspunkt der *kausalen Inferenz* ist eine ersuchte erwartungstreue und konsistente Schätzung von $E(Y_t)$ und $E(Y_c)$, die den ACE definieren. Schätzen lässt sich der Erwartungswert $E(Y_t)$ ausgehend von beobachteten Daten $(y_i, s_i, \vec{x}_i)$ lediglich für die Einheiten, für die $S_i = 1$, da entsprechend der SUTVA die beobachteten Daten übereinstimmen mit $(y_{i,t}, s_i, \vec{x}_i)$; für die Einheiten mit $S_i = 0$ sind entsprechende Daten *missed* und $E(Y_t)$ ist anhand der Daten weder identifiziert noch schätzbar (*Fundamentalproblem der kausalen Inferenz*). Gegeben dem Modell, dass das treatment-assignment vollständig durch $\vec{x}_i$ erklärt wird (vgl. Abschnitt 1.8.2), ist die Wahrscheinlichkeit für einen *complete case*, d.h. für die Beobachtung der Daten $(y_{i,t}, s_i, \vec{x})$, deckungsgleich mit dem propensity score, $e(\vec{x}_i)$. (43) entsprechend wird durch $\frac{S_i \cdot Y_i}{e(\vec{x}_i)}$ jede Beobachtung $Y_i = y_i$ als *complete case* bei der Schätzung von $E(Y_t)$ gewogen; analog werden die Beobachtungen $Y_i = y_i$ mit selbigen $e(\vec{x}_i)$, für die $S_i = 0$, durch $\frac{(1-S_i) \cdot Y_i}{(1-e(\vec{x}_i))}$ als *missed case* bei der Schätzung gewogen. Stuart (2010, S.19) betont:

> „This weighting serves to weight both the treated and control groups up to the full sample.“

Robins, Rotnitzky und Zhao (1994) zeigen theoretisch auf, dass sämtliche vorgeschlagenen *IPW*-Schätzer konsistent sind, insofern die Wahrscheinlichkeit für einen complete case, d.h. $e(\vec{x}_i)$, korrekt modelliert worden ist. Lunceford und Davidian (2004) zeigen empirisch in mehreren Monte-Carlo-Studien, dass die weighting-Methoden in Stichproben nur geringe Verzerrungen bei der Schätzung des ACE aufweisen und den anderen Adjustierungsverfahren hinsichtlich des resultierenden bias überlegen sind.

Insbesondere in kleinen Stichproben führt die Verwendung von weighting-Methoden zu einem möglichen Problem, das die Effizienz der Schätzer betrifft: In Folge kleiner werdender *complete case*-Wahrscheinlichkeiten gilt im Grenzwert für das verwendete Gewicht

$$\lim_{e(\vec{x}_i) \to 0} \quad \frac{1}{e(\vec{x}_i)} = \infty,$$

analog bei größer werdender complete case-Wahrscheinlichkeit:

$$\lim_{e(\vec{x}_i) \to 1} \quad \frac{1}{e(\vec{x}_i)} = -\infty,$$

so dass die ermittelten Gewichte entsprechend numerisch instabil werden können. Dem entsprechend fallen die vorgeschlagenen weighting-Schätzer explizit in kleinen Stichproben nicht

notwendigerweise in die Klasse der effizienten Schätzer, da in Folge der Instabilitäten die geschätzten Varianzen drastisch zunehmen (z.B., Lunceford & Davidian, 2004; Tan, 2010). In diesem Zusammenhang definierte Tan (2010) weitere $IPW-$Schätzer, die durch eine *double robustness* charakterisiert sind: insofern das propensity score-Modell korrekt spezifiziert ist, jedoch das Regressionsmodell, innerhalb dessen Y_i auf S_i und die Kovariablen $\mathbf{X}$ regressiert wird, fehlspezifiziert ist, zeigen sich die vorgeschlagenen Schätzer als konsistent; vice versa bei einem fehlspezifierten $e(\vec{x}_i)$ und einem korrekt spezifizierten Modell für Y_i. Zusätzlich zeichnen sich alle Schätzer dadurch aus, dass die zugehörigen Varianzen *bounded* sind, d.h., dass eine numerische Instabilität einzelner Gewichte entsprechend nicht die geschätzten Varianzen des ACE drastisch beeinflussen. Frölich (2004) weist einen weiteren Umgang mit dem Problem der numerischen Instabilität der Gewichte auf: Es wird ein sog. *trimming* vorgenommen, bei dem - vereinfacht - die statistischen Einheiten mit dem größten Gewicht bei der Analyse ausgeschlossen werden. Es wird jedoch an dieser Stelle argumentiert, dass die geschätzten Standardfehler der IPW-Schätzer abnehmen, jedoch die Schätzer explizit in kleinen Stichproben nicht mehr erwartungstreu sind.

Lunceford und Davidian (2004) definieren als Varianzschätzer für den IPW_1-Schätzer:

$$\hat{\sigma}^2_{IPW_1} = \frac{1}{n^2}\sum_{i=1}^{n}\left(\frac{S_i \cdot Y_i}{\hat{e}(\vec{x}_i)} - \frac{(1-S_i)\cdot Y_i}{(1-\hat{e}(\vec{x}_i))} - IPW_1 - (S_i - \hat{e}(\vec{x}_i))\hat{\mathbf{H}}^T_{\beta,1}\hat{\mathbf{E}}^{-1}_{\beta\beta}\mathbf{W}_i\right)^2, \tag{45}$$

wobei $\mathbf{W}_i$ der Designmatrix des propensity score-Modells entspricht und

$$\hat{\mathbf{H}}_{\beta,1} = \frac{1}{n}\sum_{i=1}^{n}\left[\frac{S_i \cdot Y_i \cdot (1-\hat{e}(\vec{x}_i))}{\hat{e}(\vec{x}_i)} + \frac{(1-S_i)\cdot Y_i \cdot \hat{e}(\vec{x}_i)}{(1-\hat{e}(\vec{x}_i))}\right]\cdot \mathbf{W}_i$$

$$\hat{\mathbf{E}}^{-1}_{\beta\beta} = \frac{1}{n}\sum_{i=1}^{n}\hat{e}(\vec{x}_i)(1-\hat{e}(\vec{x}_i))\mathbf{W}_i\mathbf{W}_i^t.$$

Der Varianzschätzer des IPW_2-Schätzers ist definiert als (Lunceford & Davidian, 2004):

$$\hat{\sigma}^2_{IPW_2} = \frac{1}{n^2}\sum_{i=1}^{n}\left(\frac{S_i(Y_i - \hat{E}(Y_{t,IPW_2}))}{\hat{e}(\vec{x}_i)} - \frac{(1-S_i)(Y_i - \hat{E}(Y_{c,IPW_2}))}{(1-\hat{e}(\vec{x}_i))} - (S_i - \hat{e}(\vec{x}_i))\hat{\mathbf{H}}^T_{\beta,2}\hat{\mathbf{E}}^{-1}_{\beta\beta}\mathbf{W}_i\right)^2, \tag{46}$$

wobei:

$$\hat{\mathbf{H}}_{\beta,2} = \frac{1}{n}\sum_{i=1}^{n}\left[\frac{S_i(Y_i - \hat{E}(Y_{t,IPW_2}))(1-\hat{e}(\vec{x}_i))}{\hat{e}(\vec{x}_i)} + \frac{(1-S_i)Y_i\hat{e}(\vec{x}_i)}{(1-\vec{x}_i)}\right]\cdot \mathbf{W}_i$$

2 Fragestellungen

In dem Kapitel 1.11 wurden die drei propensity score-Methoden, die für eine adjustierte Schätzung des ACE in observational studies vorliegen, hergeleitet und die Eigenschaften, die diese Adjustierungsmethoden bei der Schätzung des interessierenden treatment-Effektes besitzen, in theoretischer Hinsicht aufgeführt.

Es lassen sich wenige Veröffentlichungen finden, in denen ein direkter empirischer Vergleich der verschiedenen propensity score-Methoden bei der Schätzung des treatment-Effekts dargeboten wird; zumeist handelt es sich bei diesen Veröffentlichungen um den empirischen Vergleich eines neuen Adjustierungsvorschlags mit bestehenden propensity score-Methoden, der dargelegt wird, um die theoretisch hergeleiteten Eigenschaften des vorgeschlagenen Schätzers anhand von Simulationsstudien zu bekräftigen. Als Beispiele hierfür lassen sich der Vergleich vorgeschlagener propensity score-weighting-Methoden mit den konventionellen propensity score-weighting-Verfahren (Tan, 2010) oder der Vergleich eines neuen Vorschlags für ein propensity score-matching mit bestehenden propensity score-matching-Verfahren nennen (z.B., Abadie & Imbens, 2006; Dehejia & Wahba, 2002). Zwar lassen sich auch einige systematische Vergleiche diverser propensity score-Verfahren untereinander finden, diese fokussieren aber zumeist nur eine Auswahl an verschiedenen Adjustierungsmethoden: Lunceford und Davidian (2004) gleichen beispielsweise die Ergebnisse der propensity score-Stratifikation mit den Ergebnissen der beiden in Kapitel 1.11 aufgeführten Verfahren des propensity score-weightings in mehreren Simulationsszenarien bei der Schätzung des treatment-Effekts ab, Morgan und Winship (2007) führten einen systematischen Vergleich mehrerer propensity score-matching-Verfahren auf.

Damit erscheint es naheliegend, in mehreren Simulationsszenarien die Güte aller drei propensity score-Methoden bei der Schätzung des ACE empirisch zu untersuchen und somit einen systematischen Vergleich aller vorliegenden Methoden berichten zu können. In dem ersten Teil dieser Arbeit werden die Ergebnisse von insgesamt fünf Simulationsszenarien aufgeführt, in denen zwei verschiedene Stratifikationsverfahren, die beiden aufgeführten weighting-Methoden sowie das exakte matching, sämtlich auf dem propensity score basierend, mit einander verglichen werden sollen. Zu diesem Zweck werden auch Simulationsszenarien generiert, in denen das logistische Regressionsmodell zur Schätzung des unbekannten propensity scores von dem datengenerierenden Prozess abweicht und somit eine Fehlspezifikation des modellierten treatment-assignments aufweist. Durch dieses Vorgehen lassen sich entsprechende Konsequenzen, die eine

fehlspezifizierte propensity score-Schätzung hat, bei der Schätzung des ACE aufweisen und mögliche Unterschiede zwischen den verschiedenen Adjustierungsmethoden hinsichtlich resultierender Verzerrungen oder zunehmender Standardfehler berichten.

Ausgehend von der Tatsache, dass der propensity score der Einheit i ein unbekannter Skalar ist und für eine adjustierte Schätzung des ACE entsprechend geschätzt werden muss, wird im zweiten Teil dieser Arbeit ein non-parametrisches Schätzverfahren für den unbekannten propensity score vorgeschlagen, das sich in dem speziellen Szenario, in dem die konfundierenden Störvariablen der Klasse der stetigen Kovariablen mit log-konkaven Dichten zugehören, anwenden lässt und auf eine Modellierung des propensity scores, beispielsweise in einem logistischen Regressionsmodell, verzichtet. Hierzu wird zunächst die Definition des propensity scores als bedingte Wahrscheinlichkeit für eine Umformulierung über den Satz von Bayes genutzt, um zu zeigen, dass die resultierende bedingte Wahrscheinlichkeit lediglich von den unbekannten, am Selektionsindikator bedingten Dichtefunktionen der stetigen Kovariablen abhängig ist. Es wird gezeigt, dass sich diese unbekannten bedingten Dichten mit einem Maximum Likelihood-Schätzer, der von Cule et al. (2010) für die Klasse der log-konkaven Dichten definiert worden ist, schätzen lassen und mit dieser Schätzung folgend der propensity score der Einheit i identifiziert ist. Nach theoretischer Herleitung dieses Schätzverfahrens und Vorstellung des verwendeten ML-Schätzers werden in mehreren Simulationsszenarien die Eigenschaften dieses Schätzverfahrens hinsichtlich einer adjustierten Schätzung des treatment-Effekts evaluiert. Unter Anderem werden in diesen Szenarien das Vorliegen einer univariaten, einer multivariaten und einer heavy-tailed log-konkaven Dichte sowie das Vorliegen einer non-log-konkaven Dichtefunktion simuliert, die damit einen Rückschluss auf die Güte der Schätzungen in verschiedenen für die Praxis relevante Szenarien und auf mögliche Annahmeverletzungen, die der ML-Dichteschätzung zu Grunde liegen, ermöglichen.

In theoretischer Hinsicht wird in der Annahme des strongly ignorable treatment-assignment sowie in der daraus abgeleiteten Definition des propensity scores eine messfehlerfreie Erhebung der Kovariablen $\vec{x}_i$ impliziert, so dass damit manifeste Störvariablen unterstellt werden. In dem dritten Teil dieser Arbeit soll auf das Problem der latenten Variablen eingegangen werden, dem explizit die Psychologie als empirische Humanwissenschaft ausgesetzt ist. An dieser Stelle wird von dem Szenario ausgegangen, dass die konfundierende Störvariable latent, d.h. empirisch nicht direkt zugänglich sondern nur messfehlerbelastet durch Indikatorvariablen operationalisiert,

vorliegt. Zunächst wird in diesem Teil der Arbeit durch die Vorstellung des errors-in-variables-Problems die Konsequenz aufgeführt, die eine Schätzung des treatment-Effekts, adjustiert an den beobachteten Indikatorvariablen anstelle der konfundierenden latenten Störvariablen, hat; davon ausgehend wird als eine gängige parametrische Methode zur adjustierten Schätzung des treatment-Effekts die Strukturgleichungsmodellierung vorgestellt, die eine Schätzung des interessierenden Effekts unter Konstanthaltung der Messfehler, die sich bei der Operationalisierung der latenten Variablen durch die Indikatorvariablen einstellen, ermöglicht. Der Strukturgleichungsmodellierung wird ein Vorschlag zur Adjustierung gegenübergestellt, der auf den resultierenden Hauptkomponenten der zu den Indikatorvariablen gehörigen Korrelationsmatrix beruht und die extrahierten Hauptkomponenten folgend als Prädiktorvariablen in das logistische Regressionsmodell zur Schätzung der unbekannten propensity scores aufnimmt. Nach Herleitung dieser Adjustierungsmethode durch Vorstellung der Hauptkomponentenanalyse und der Zusammenführung dieser mit der Definition des propensity scores werden in mehreren Simulationsszenarien die Ergebnisse der adjustierten Schätzung des treatment-Effekts, einerseits nach Adjustierung gegeben der Strukturgleichungsmodellierung, andererseits nach Adjustierung mit dem vorgeschlagenen propensity score, aufgeführt. Zu diesen Simulationsszenarien zählen auch mögliche Szenarien, in denen die Annahmen, die der Strukturgleichungsmodellierung als parametrtisches Verfahren zu Grunde liegen, verletzt werden und damit mögliche Konsequenzen, die sich in Folge dessen einstellen, aufgewiesen werden können.

3 Vergleich verschiedener Propensity Score-Methoden

3.1 Einleitung und Methodik

Im Folgenden werden die Ergebnisse mehrerer Simulationsszenarien vorgestellt, die alle mit der Zielsetzung generiert worden sind, die vorherig aufgeführten propensity score-Methoden hinsichtlich ihrer Erwartungstreue, Konsistenz und Effizienz bei der Schätzung des ACE zu untersuchen und zu vergleichen. In den folgenden Simulationen wird das Verhalten dieser Adjustierungsmethoden unter variierenden Bedingungen überprüft: So werden die propensity score-Methoden bei Vorliegen einer diskreten Störvariablen (Simulation 1), bei Vorliegen einer stetigen Störvariablen (Simulation 2) sowie im Falle mehrerer Fehlspezifikationen des logistischen Regressionsmodells, das zur Schätzung des propensity scores spezifiziert wird (Simulation 3, ff.), hinsichtlich der Schätzung des ACE evaluiert.

Im Vergleich zu veröffentlichten Simulationsstudien (z.B., Dehejia & Wahba, 2002; Drake, 1993; Zhao, 2004), werden im Folgenden mehrere Adjustierungsmethoden gleichzeitig analysiert, so dass die Ergebnisse verschiedener Methoden innerhalb einer Simulation ins Verhältnis zu einander gesetzt werden können. Die verwendeten propensity score-Methoden, die in den folgenden fünf Simulationen begutachtet werden, sind

a) die Dezentil-Stratifikation

b) die Quintil-Stratifikation

c) der IPW_1-Schätzer,

d) der IPW_2-Schätzer,

e) das exakte propensity score-matching.

3.1.1 Aufbau der Simulationsstudien

Es wurde für alle Simulationsstudien eine finite Grundgesamtheit mit $N = 10.000$ statistischen Einheiten definiert, die gemäß dem Rubin-Modell durch die zwei potentiellen response-Werte $y_{i,t}$ und $y_{i,c}$ charakterisiert sind, wobei $y_{i,t}$ den response-Wert unterhalb von t indiziert, den i annimmt, wenn i nach Stichprobenziehung zu t selegiert wird - vice verca für $y_{i,c}$. Für die finite Grundgesamtheit wurden die potentiellen response-Werte durch $Y_c \sim \mathcal{N}(100, 15^2)$ und $Y_t = Y_c$

realisiert, so dass: ACE $= E(Y_t) - E(Y_c) = 0$ - es handelt sich damit um den zu replizierenden Effekt.

Der faktisch beobachtete response, y_i, wurde durch ein einfaches lineares Modell spezifiziert, innerhalb dessen sich der beobachtete response additiv aus dem jeweiligen potentiellen response-Wert der Bedingung, zu der i selegiert wird, und bestimmten Kovariablenwerten zusammensetzt; mit ACE $= 0$ gilt: $y_i = y_{i,c} + \vec{x}_i^t \vec{\beta}$, wobei $\vec{x}_i$ und $\vec{\beta}$ je nach Simulationsstudie variieren. Für die ersten beiden Simulationsszenarien wurde die Kovariablenmatrix $\mathbf{X}$ definiert durch die Realisierungen folgender Verteilungsmodelle:

- $X_1 \sim B(1, 0.5)$
- $X_2 \sim \mathcal{N}(30, 5)$
- $X_3 = \gamma_1 \cdot X_2 + \epsilon_i$, $\gamma_1 = 1$, $\epsilon_i \sim \mathcal{N}(0, 1)$ und $\epsilon_i \perp\!\!\!\perp X_2$
- $X_4 \sim \mathcal{U}\left(\frac{1}{4}\right)$, $\mathcal{X}_4 := \{1, 2, 3, 4\}$
- $X_5 = \gamma_2 \cdot X_3 + \epsilon_i$, $\gamma_2 = 1$, $\epsilon_i \sim \mathcal{N}(0, 20)$ und $\epsilon_i \perp\!\!\!\perp X_3$
- $X_6 \sim \mathcal{N}(2, 10)$

Für diejenigen Simulationen, innerhalb derer eine Fehlspezifikation des propensity score-Modells von Interesse war (Simulation 3, ff.), bestand die Kovariablenmatrix $\mathbf{X}$ aus den Realisierungen des Vektors $\vec{X}^t = (X_1, ..., X_5)$, wobei $\vec{X} \sim \mathcal{N}(\vec{0}, \mathbf{\Sigma})$ mit der positiv-definiten Varianz-Kovarianzmatrix

$$\mathbf{\Sigma} = \begin{pmatrix} 1 & & & & \\ 0.2 & 1 & & & \\ 0.1 & 0.2 & 1 & & \\ 0.4 & 0.3 & 0.1 & 1 & \\ 0.2 & 0.1 & 0.3 & 0.2 & 1 \end{pmatrix}.$$

Das entsprechende response-Modell wird zu Beginn einer jeden Simulation konkretisiert.

Es werden in jeder Wiederholung eines Simulationsszenarios jeweilig $n \in \{100, 200, 500\}$ Einheiten als einfache Zufallsstichprobe gezogen, wobei die jeweilige Simulation für jeden Stichprobenumfang $k = 5000$fach wiederholt wird. Nach Stichprobenziehung setzt das treatment-assignment der Stichprobeneinheiten ein, wobei für alle Simulationsstudien ein logistisches

(Schwellenwert-)Modell mit

$$\Pr(S_i = 1 \mid \vec{x}_i) = \frac{1}{1 + \exp(-(\alpha_0 + \vec{x}_i^t \vec{\alpha}))} \qquad \alpha_0 = 0$$

zu Grunde gelegt worden ist. Mit $\alpha_0 = 0$ werden die Chancen, zu t selegiert zu werden, vollständig durch $\vec{x}_i$ determiniert. $\vec{\alpha}$ sowie $\vec{x}_i$ werden in den Einleitungen der fünf Simulationsszenarien spezifiziert.

In jeder gezogenen Stichprobe wurde für eine naïve Punktschätzung des ACE die Stichprobenfunktion $\bar{Y}_t - \bar{Y}_c$ gewählt, um so eine empirische Evidenz über den bias zu bekommen, der resultiert, wenn die Schätzung nicht durch mögliche propensity score-Methoden adjustiert wird. Hierzu wurden nach den $k = 5000$ Wiederholungen pro n die resultierenden Punktschätzungen gemittelt - dieses arithmetische Mittel dient folgend als empirisches Indiz für $E\left(\bar{Y}_t - \bar{Y}_c\right)$.

Der unadjustierten Schätzung des ACE folgend wurde in jeder gezogenen Stichprobe der propensity score für die Stichprobeneinheiten mittels einer logistischen Regression geschätzt; die *fitted values*, d.h. die an $\vec{x}_i$ bedingten, geschätzten Modellwahrscheinlichkeiten, zu t selegiert zu werden, wurden als geschätzte propensity scores, $\hat{e}(\vec{x}_i)$, behandelt. Die Spezifikation des linearen Prädiktors η_i, der zur Schätzung von $e(\vec{x}_i)$ modelliert wurde, wird zu Beginn jeder der folgenden Szenarien aufgeführt. Die geschätzten propensity scores wurden daraufhin für die oben aufgeführten Adjustierungsmethoden zur Schätzung des ACE verwandt (folgend allgemein: $\hat{\tau}_{\text{adjustiert}}$):

Die Dezentil-Stratifikation wurde durchgeführt, indem $v = 10$ strata, basierend auf den Dezentilen der $\hat{e}(\vec{x}_i)$, definiert worden sind. Innerhalb eines jeden stratums wird der stratumsspezifische ACE durch $\hat{\tau}_j = \bar{Y}_{t,j} - \bar{Y}_{c,j}, j = 1, \ldots, 10$, geschätzt. Insofern Extremstrata auftraten, innerhalb derer die Einheiten vollständig einer der Bedingungen zugeordnet waren, wurden diese von der Analyse ausgeschlossen, da eine Schätzung des ACE hier nicht möglich war. Der stratumsunspezifische ACE wird, wie in (39) aufgeführt, durch eine gewogene Mittelung der schichtspezifischen Punktschätzer $\hat{\tau}_j$ geschätzt.

In einem stratum j erfolgte die schichtspezifische Varianzschätzung des ACE, $\hat{\sigma}^2_{\hat{\tau}_j}$, gemäß (40), wobei mit Definition identischer Varianzen der response-Werte beider Gruppen, d.h. $\sigma_t^2 = \sigma_c^2 = 15^2$ (s. oben), in jedem stratum die Stichprobenvarianzen aus t und c gepooled werden konnten, so dass:

$$\hat{\sigma}^2_{\hat{\tau}_j} = \left(\frac{1}{n_{t,j}} + \frac{1}{n_{c,j}}\right) \cdot \frac{(n_{t,j} - 1) \cdot \hat{\sigma}^2_{t,j} + (n_{c,j} - 1) \cdot \hat{\sigma}^2_{c,j}}{n_{t,j} + n_{c,j} - 2}$$

Die allgemeine Varianzschätzung des ACE, folgend für Intervallschätzungen des ACE benötigt, erfolgte durch eine gewogene Mittelung der schichtspezifischen Varianzen, wie in (41) aufgeführt.

Um aufweisen, wie stark die Güte der Schätzung des ACE von der Anzahl an Quantilen abhängig ist, wurde eine, im Verhältnis zur Dezentil-Stratifikation grobmaschigere, Quintil-Stratifikation gegeben der geschätzten propensity scores durchgeführt. Im Unterschied zur Dezentil-Stratifikation wurden $v = 5$ strata definiert; sämtliches weiteres Vorgehen entspricht dem der Dezentil-Stratifikation.

Es wurden nebst Stratifikationsmethoden beide aufgeführten weighting-Methoden, IPW_1 und IPW_2, zur Schätzung des ACE verwandt. Die geschätzten Varianzen des ACE ergaben sich unter Verwendung der in (45) und (46) definierten Varianzschätzer. Es wurden entgegen der Vorschläge zur Varianzstabilisierung keine weitere Adjustierungen bezüglich möglicherweise auftretender Extremgewichte, beispielsweise durch ein trimming, vorgenommen, da für eine generelle Einschätzung der Güte als Schätzer Szenarien, in denen diese Extremwerte auftreten können, von Interesse sind und entsprechend evaluiert werden sollen.

Als dasjenige matching-Verfahren, das den beiden anderen propensity score-Methoden gegenüber gestellt werden soll, wurde das *exakte matching* gewählt. Das exakte matching wurde durchgeführt, indem zwei Einheiten gematched wurden, die den selben geschätzten propensity score besitzen. Diejenigen Einheiten, die nicht gematched werden können, werden von der anschließenden Schätzung des ATT ausgeschlossen. Das exakte matching als eines der möglichen matching-Verfahren wurde für diese Simulationsstudien gewählt, da es in theoretischer Hinsicht stets als Idealvorstellung eines matching-Verfahren gilt und ein approximatives matching bereits durch die beiden Stratifikationsverfahren abgedeckt ist.

3.1.2 Gütekriterien

Die Güte der Schätzungen eines bestimmten Adjustierungsverfahrens, $\hat{\tau}_{adjustiert}$, wird folgend beurteilt

a) durch eine Schätzung der Erwartungstreue. Hierzu wurden im Anschluss der $k = 5000$ Wiederholungen pro n und für jedes der Adjustierungsverfahren die Punktschätzungen gemittelt, um so empirische Evidenz über die Erwartungstreue zu erhalten. Gegeben dem Mittel der Punktschätzer soll geprüft werden, ob $m\left(\hat{\tau}_{\text{adjustiert}}\right) = 0$. Durch ACE $= 0$ bietet sich der Vorteil, dass das resultierende Mittel der Punktschätzungen eine direkte

Schätzung für den bias der Schätzstatistik ermöglicht.

b) durch eine empirische Einschätzung der M.S.E.-Konsistenz. Hierzu wird für jeden der Stichprobenumfänge der *mean squared error* (folgend: M.S.E.) geschätzt, wobei:

$$\text{M.S.E.} = \text{bias}^2_{\hat{\tau}_{\text{adjustiert}}} + \text{Var}(\hat{\tau}_{\text{adjustiert}}).$$

Für einen (asymptotisch) erwartungstreuen Schätzer sollte sich der geschätzte M.S.E. mit größer werdenden n null näheren.

c) durch eine Schätzung der Effizienz. Hierzu wird für jedes der Adjustierungsverfahren der zugehörige Standardfehler durch die empirische Standardabweichung der Punktschätzer geschätzt. Zusätzlich werden die ranges der Punktschätzungen berichtet.

d) durch eine empirische Bestimmung der prozentualen bias-Reduktion, die folgend definiert ist als

$$\text{bias-Reduktion} = \frac{\text{bias}_{\text{adjustiert}} - \text{bias}_{\text{unadjustiert}}}{\text{bias}_{\text{unadjustiert}}}$$

e) durch die Ermittlung der coverage als der prozentuale Anteil an 95%−igen Stichprobenkonfidenzintervalle, die den ACE überdecken. Zur Ermittlung der Konfidenzintervalle wird folgend eine Normalverteilung der Punktschätzer zu Grunde gelegt, da a) die potentiellen response-Werte als Realisierungen einer Normalverteilung vorliegen und b) die gewählten Stichprobenumfänge groß genug erscheinen, um eine entsprechende Verteilungsannahme der Schätzstatistiken zu rechtfertigen.

3.2 Simulation 1: Diskrete Störvariable

In dem folgenden Simulationsszenario sollen die aufgeführten propensity score-Methoden hinsichtlich der oben genannten Kriterien evaluiert werden, wobei das einfachste Szenario zugrunde gelegt wurde: es liegt eine binäre Variable X_1 vor, d.h.: $\mathcal{X}_1 := \{0, 1\}$, die einerseits das treatment-assignment bedingt, andererseits den response-Wert von i determiniert. Für dieses Szenario wurden folgende Modelle festgelegt, um X_1 als Störvariable zu klassifizieren:

$$y_i = y_{i,c} + \beta_1 \cdot x_{i,1}, \quad \beta_1 = 10,$$

sowie:

$$\Pr(S_i = 1 \mid \vec{x}_i) = \frac{1}{1 + \exp(-(\ln(2) \cdot x_{1,i}))}.$$

Damit haben statistische Einheiten mit $x_{1,i} = 1$ eine zweifach größere Chance, zu t selegiert zu werden. Die Schätzung des propensity scores in einer Stichprobe erfolgte mittels einer logistischen Regression, innerhalb derer sämtliche Kovariablen, $\vec{X}^t = (X_1, \ldots, X_6)$, in einfacher Form zur Modellspezifikation benutzt wurden.

Die Ergebnisse sind folgend tabellarisch sowie, wenn für die Befundlage relevant, graphisch in Form von 95%−igen Stichprobenkonfidenzintervallen dargestellt.

Ergebnisse

Tabelle 1 führt die Ergebnisse der Simulation auf: Es ist in Anbetracht der Ergebnisse erkenntlich, dass sämtliche propensity score-Methoden in der Lage sind, den bias, der einer unadjustierten Schätzung des ACE folgt, zu reduzieren (s. Tabelle 1). Sämtliche Adjustierungsverfahren weisen durchschnittliche Schätzungen auf, die nur geringfügig von dem ACE abweichen. Es lassen sich bereits in kleinen Stichproben auffällige bias-Reduktionen festhalten, die damit aufweisen, dass mit einer propensity score-Adjustierung, unabhängig von der genutzten Methode, die konfundierende Wirkung der diskreten Störvariablen minimiert werden kann. Einhergehend mit den theoretischen Überlegungen zeigt sich, dass explizit in kleinen Stichproben einerseits das exakte matching, andererseits die Qunitil-Stratifikation die verhältnismäßig größte Verzerrung bzw. die niedrigste bias-Reduktion aufweisen - ein Befund, der sich beim matching unmittelbar auf die Stichprobengröße zurückführen lässt, da in einer kleinen Stichprobe die Wahrscheinlichkeit, einen exakten match zu finden, geringer ist als in großen Stichproben. So zeigt sich, dass mit $n = 500$ das exakte matching-Verfahren dasjenige Verfahren ist, für das in

Folge der größeren matching-Wahrscheinlichkeit die größte bias-Reduktion resultiert. Es lässt sich zusätzlich für das exakte matching eine Reduktion des Standardfehlers mit wachsenden Stichprobenumfang festhalten, so dass die resultierende M.S.E.-Schätzung mit zunehmenden Stichproben abnimmt.

Tabelle 1: Ergebnisse der Simulation 1

	Methode	*bias*	*Reduk.*	*S.E.*	*M.S.E.*	*Cov.*	*range*
$n = 100$	unadj.	1.69	-	3.17	12.91	87.28	23.17 (-9.72, 13.44)
	Dez.-Strat.	0.05	97.34	3.47	12.04	95.08	25.17 (-12.33, 12.83)
	Quint.-Strat.	0.11	93.49	6.52	42.52	93.28	51.60 (-26.44, 25.15)
	IPW_1	0.06	96.45	4.91	24.11	97.86	74.53 (-43.92, 30.60)
	IPW_2	0.04	97.63	3.19	10.18	93.02	23.55 (-11.46, 12.08)
	ex. mat.	0.10	94.08	4.23	17.90	94.02	32.87 (-16.62, 16.25)
$n = 200$	unadj.	1.69	-	2.24	7.87	81.50	16.03 (-6.46, 9.57)
	Dez.-Strat.	0.02	98.93	2.31	5.34	95.68	16.16 (-8.21, 7.94)
	Quint.- Strat.	0.05	96.98	4.74	22.47	93.30	35.73 (-17.73, 18.00)
	IPW_1	-0.005	99.70	2.67	7.13	96.94	25.14 (-11.92, 13.22)
	IPW_2	-0.016	99.05	2.15	4.62	94.70	16.41 (-8.33, 8.075)
	ex. mat.	0.001	99.93	2.93	8.58	95.02	20.58 (-9.95, 10.62)
$n = 500$	unadj.	1.68	-	1.42	4.83	68.66	10.87 (-3.79, 7.08)
	Dez.-Strat.	0.01	98.75	1.40	1.97	95.08	9.58 (-4.91, 4.66)
	Quint.-Strat.	0.04	97.74	2.88	8.30	92.04	22.34 (-11.17, 11.17)
	IPW_1	-0.015	99.10	1.47	2.16	95.24	9.96 (-4.85, 5.10)
	IPW_2	-0.017	98.98	1.39	1.92	94.60	9.74 (-4.82, 4.91)
	ex. mat.	-0.01	99.42	1.81	3.26	95.28	13.63 (-7.2, 6.42)

Anm.: ACE = 0, *Reduk.:* prozentuale bias-Reduktion, *Cov.:* prozentuale Coverage, unadj.: unadjustierte Schätzung des ACE, Dez.-Strat.: Dezentil-Stratifikation, Quint.-Strat.: Quintil-Stratifikation, IPW_1 : inverse probability weighting 1, IPW_2 : inverse probability weighting 2, ex. mat.: exaktes matching

Die Konfidenzintervalle werden im Allgemeinen dergestalt geschätzt, dass die coverage über

alle Adjustierungsverfahren hinweg mit dem festgelegten Vertrauensniveau von 95% vereinbar ist; einzige Ausnahme stellt hier das Verfahren der Quintil-Stratifikation dar, welches unter allen Adjustierungsverfahren dasjenige Verfahren mit dem größten resultierenden bias, Standardfehler und M.S.E. ist. Damit erkenntlich ist, dass in Abhängigkeit von der Anzahl an Quantilen, die zur Stratifikation genutzt wird, der bias, der geschätzte Standardfehler sowie die coverage variieren - ein Befund, der zurückzuführen ist auf eine größere Varianz der geschätzten propensity scores innerhalb eines stratums bei Nutzung der Quintil-Stratifikation im Verhältnis zur Dezentil-Stratifikation. Ausgehend von einer Quintil-Stratifikation liegen die Punktschätzungen für alle n in einem größeren range und der geschätzte Standardfehler weist im Verhältnis zu allen anderen Adjustierungsverfahren die größte Unsicherheit bei der Schätzung des ACE auf. Eine graphische Inspektion der zur Dezentil- und Quintil-Stratifikation gehörigen Konfidenzintervalle bei $n = 100$ verdeutlicht diesen Befund eindeutig (vgl. Abb.2).

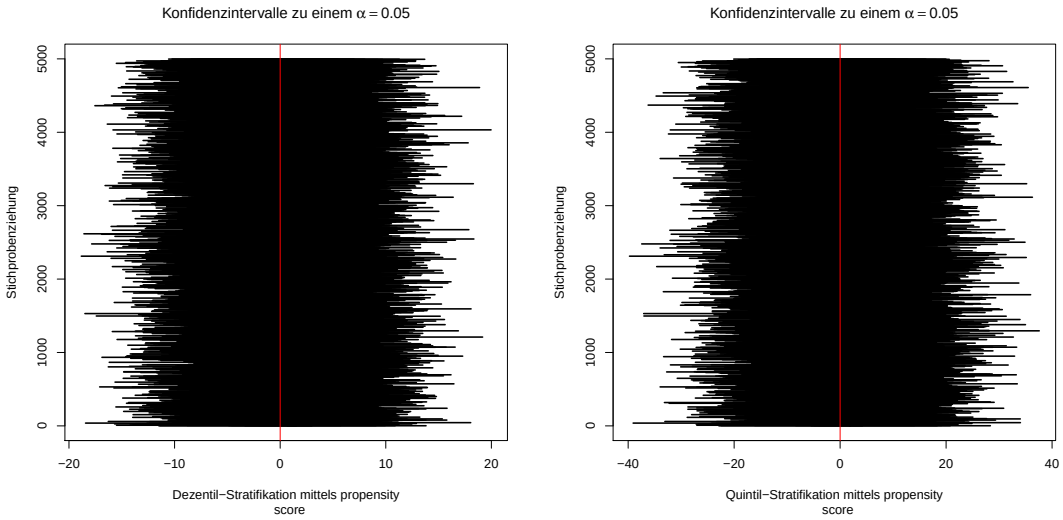

Abbildung 2: 95%−Stichprobenkonfidenzintervalle der Stratifikationsmethoden, $n = 100$

Hervorzuheben sind die Schätzergebnisse der beiden weighting-Methoden: Mit $n = 100$ weist der IPW_1-Schätzer die größten Extrema bei Punktschätzung des ACE auf. Dieses Ergebnis lässt sich zurückführen auf den Umstand, dass in vereinzelten Stichproben extreme propensity scores beobachtet werden, die in einer numerischen Instabilität der Gewichte und einer entsprechend großen Varianzschätzung des ACE münden. Die graphische Inspektion der zugehörigen Stichprobenkonfidenzintervalle für $n = 100$ weist entsprechend einzelne extreme Intervallschätzungen auf, die auf diesen extremen Varianzschätzungen basieren (vgl. Abb. 3). Dieses Ergebnis rela-

tiviert sich mit zunehmenden Stichprobenumfang, da derartig extreme Gewichte stärker bei der Varianzschätzung gemittelt werden, wie sich in Anbetracht des resultierenden ranges der Punktschätzungen zeigt. Die Punktschätzungen des IPW_2-Schätzers weisen eine solche Tendenz nicht auf: Bereits in kleinen Stichproben handelt es sich um denjenigen Schätzer mit der geringsten Variabilität und der größten bias-Reduktion.

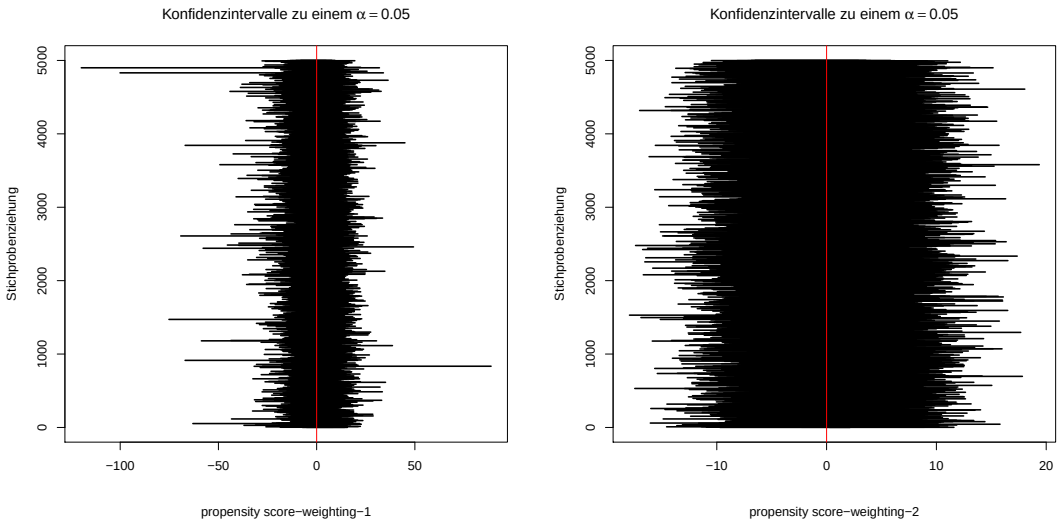

Abbildung 3: 95%−Stichprobenkonfidenzintervalle der weighting-Methoden

3.3 Simulation 2: Stetige Störvariable

In folgender Simulation wurde die Kovariable X_6 als konfundierende Störvariable genutzt. Der beobachtete response der Einheit i wurde durch das additive Modell

$$y_i = y_{i,c} + \beta_1 \cdot x_{i,6} \quad \beta_1 = 10,$$

das treatment-assignment durch das logistische Modell

$$\Pr(S_i = 1 \mid \vec{x}_i) = \frac{1}{1 + \exp(-(\ln(1.25) \cdot x_{i,6}))}$$

realisiert. Abbildung 4 veranschaulicht das treatment-assignment in Abhängigkeit von der Kovariablen; hier dargestellt ist das realisierte treatment-assignment einer konkreten Stichprobenziehung. Der propensity score in einer Stichprobe wurde analog zur ersten Simulation mit allen zur Verfügung stehenden Kovariablen geschätzt.

Ergebnisse

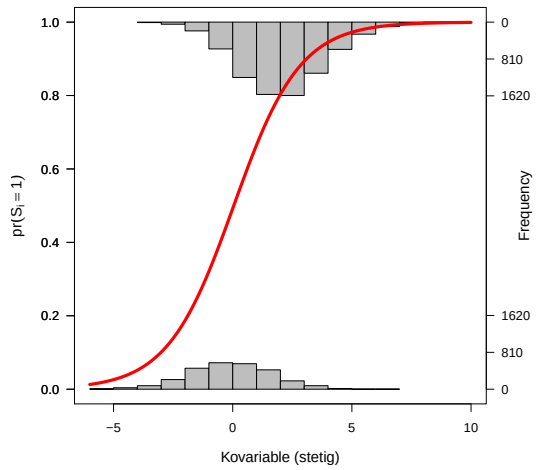

Abbildung 4: treatment-assignment in Abhängigkeit von X_6

Tabelle 2 führt die Ergebnisse der Simulationen auf: Grundsätzlich lässt sich in Anbetracht der bias-Reduktionen festhalten, dass gegeben einer normalverteilten Störvariablen sämtliche Adjustierungsverfahren in der Lage sind, den bias, der einer unadjustierten Schätzung des ACE folgt, zu reduzieren. Jedoch nähert die Vielzahl der gemittelten Punktschätzungen erst mit $n = 500$ dem ACE an, wobei diese Tendenz insbesondere beim IPW_1-Schätzers auffällt: Während mit $n = 100$ sowohl ein auffälliger bias als auch eine große M.S.E.-Schätzung, zurückzuführen auf einen extremen range an Punktschätzungen, resultieren, nehmen diese Kennzahlen mit zunehmenden n ab (die bias-Reduktion und coverage entsprechend zu), so dass mit $n = 500$ der IPW_1-Schätzer das einzige Adjustierungsverfahren ist, das in durchschnittlichen Schätzungen nahe dem ACE resultiert. Die geschätzten Standardfehler sowie die ranges der Punktschätzungen weisen extreme Unsicherheiten beim Schätzen des ACE mit der IPW_1-Methode auf, wobei in Anbetracht des ranges eine Tendenz zu extremen, negativen Punktschätzungen auffällt. Für alle Stichprobengrößen lässt sich festhalten, dass der IPW_1-Schätzer dasjenige Adjustierungsverfahren mit dem größten geschätzten Standardfehler ist, der mit zunehmendem n abnimmt. Die 95%−igen Stichprobenkonfidenzintervalle für $n = 100$ und $n = 500$ (Abb. 5) verdeutlichen, dass dieses Ergebnis auf extremen Gewichten beruht, die in einzelnen Stichproben resultieren. Erkenntlich ist anhand der Graphiken, dass die Breite der geschätzten Intervalle mit zunehmenden Stichprobenumfang entsprechend abnimmt. Explizit in kleinen Stichproben fällt auf, dass der IPW_2-Schätzer sowie das exakte matching Probleme bei der Adjustierung einer stetigen Störvariablen haben: Beide Verfahren weisen für $n = 100$ einen großen bias auf. Beim Vergleich der coverages fällt auf, dass trotz des bestehenden bias das exakte matching Intervallschätzungen liefert, die mit einem Irrtumsniveau von $\alpha = 0.05$ vereinbar sind, während der IPW_2-Schätzer mit einer coverage von 53.72% dies nicht gewährleistet (vgl. Abb. 6).

Tabelle 2: Ergebnisse der Simulation 2

	Methode	*bias*	*Reduk.*	*S.E.*	*M.S.E.*	*Cov.*	*range*
$n = 100$	unadj.	126.85	-	15.88	16343.1	0.00	115.08 (74.28, 189.36)
	Dez.-Strat.	2.33	98.16	9.56	96.81	98.16	74.12 (-36.86, 37.26)
	Quint.-Strat.	6.17	95.14	15.27	271.24	95.46	106.69 (-45.41, 61.28)
	IPW_1	8.37	93.40	47.69	2344.34	59.58	3384.73 (-3270.06, 114.67)
	IPW_2	21.37	83.15	23.48	1007.77	53.72	303.18 (-191.1, 112.08)
	ex. mat.	19.48	84.65	19.49	759.33	94.68	158.64 (-46.61, 112.02)
$n = 200$	unadj.	126.79	-	11.21	16201.37	0.00	81.27 (87.56, 168.82)
	Dez.-Strat.	2.66	97.90	5.43	36.56	95.80	36.57 (-18.29, 18.27)
	Quint.-Strat.	7.96	93.72	10.27	168.83	89.54	77.74 (-36.07, 41.67)
	IPW_1	2.17	98.28	34.50	1194.96	63.90	2568.46 (-2475.98, 92.47)
	IPW_2	11.53	90.21	19.51	513.58	59.36	343.32 (-256.95, 86.36)
	ex. mat.	12.84	89.87	12.31	316.40	93.04	104.01 (-35.29, 68.73)
$n = 500$	unadj.	126.71	-	7.072	16105.44	0.00	58.41 (101.92, 160.33)
	Dez.-Strat.	3.74	97.05	3.77	28.20	88.20	26.31 (-8.24, 18.08)
	Quint.-Strat.	11.73	90.74	8.35	207.31	72.74	66.08 (-21.36, 44.73)
	IPW_1	0.23	99.82	26.61	708.14	70.02	1252.33 (-1187.85, 64.47)
	IPW_2	5.23	95.87	16.53	30.59	66.76	287.26 (-233.44, 53.81)
	ex. mat.	7.21	94.31	6.82	98.50	89.94	67.38 (-22.8, 44.55)

Anm.: ACE = 0, *Reduk.:* prozentuale bias-Reduktion, *Cov.:* prozentuale Coverage, unadj.: unadjustierte Schätzung des ACE, Dez.-Strat.: Dezentil-Stratifikation, Quint.-Strat.: Quintil-Stratifikation, IPW_1 : inverse probability weighting 1, IPW_2 : inverse probability weighting 2, ex. mat.: exaktes matching

Dieser Umstand lässt sich zurückführen auf einen immensen range an negativen wie positiven Punktschätzungen nach Adjustierung mit der IPW_2-Methode, der beim exakten matching im negativen Bereich nicht derartig zu beobachten ist. Für beide Verfahren gilt, dass mit zunehmenden Stichprobenumfang sowohl die resultierenden Verzerrungen als auch die Standardfehler abnehmen, jedoch die coverage des IPW_2-Schätzers weiterhin unter dem festgelegten Niveau von 95% liegt. Insbesondere in kleinen Stichproben ($n = 100$) erweist sich die Dezentil-

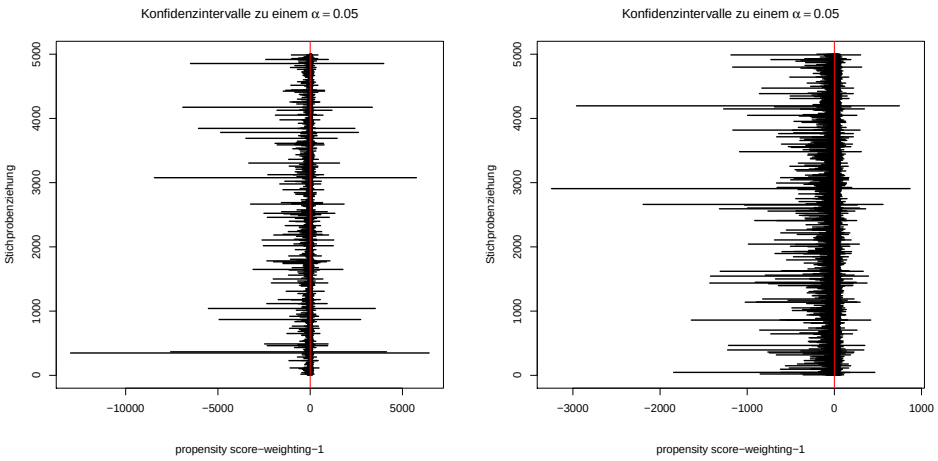

Abbildung 5: 95%−Stichprobenkonfidenzintervalle der IPW_1-Methode für $n = 100$ und $n = 500$

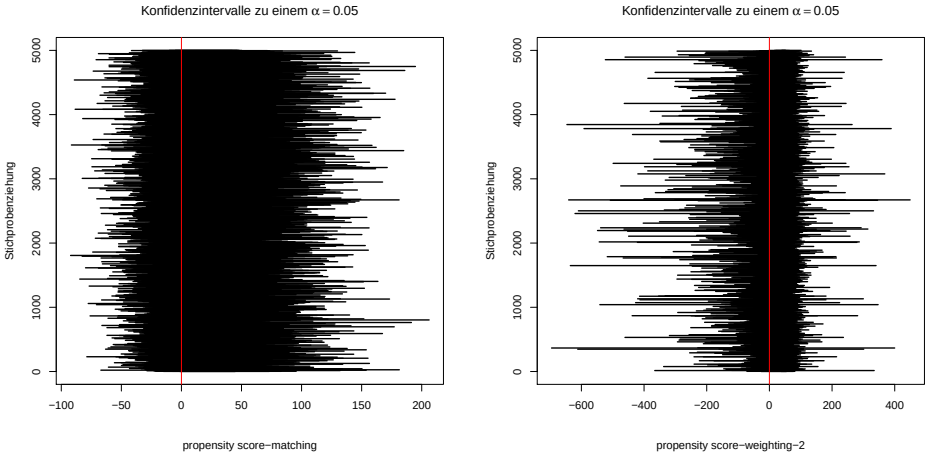

Abbildung 6: Vergleich der 95%−Stichprobenkonfidenzintervalle des matching-Verfahrens und der IPW_2-Methode für $n = 100$

Stratifikation gegenüber allen anderen Adjustierungen überlegen, da es sich hier um das Adjustierungsverfahren mit dem geringsten bias, dem kleinsten Standardfehler und der größten coverage handelt. Es zeigt sich mit zunehmenden Stichprobenumfang, dass der resultierende bias der Dezentil-Stratifikation geringfügig zunimmt - ein Umstand der sich unmittelbar erklären lässt, da mit zunehmenden n eine größere Varianz der propensity scores innerhalb eines stratums resultiert -, die geschätzten Standardfehler abnehmen und die coverage entsprechend mit abnimmt. Selbige Tendenz lässt sich für die Qunitil-Stratifikation festhalten: Mit zunehmenden n resultieren zunehmende Verzerrungen, so dass mit $n = 500$ der resultierende bias der Quintil-Stratifikation der größte aller Schätzverfahren ist.

3.4 Simulationen 3ff.: Fehlspezifizierte Propensity Score-Modelle

In Abschnitt 1.10.3 wurde aufgeführt, dass in praxi die Schätzung des unbekannten propensity scores zumeist durch das parametrische Verfahren der logistischen Regression erfolgt. Um $e(\vec{x}_i)$, folglich auch den ACE nach Adjustierung, konsistent schätzen zu können, ist es entsprechend notwendig, den linearen Prädiktor des propensity score-Modells, η_i, vollständig und korrekt zu spezifizieren. Im Folgenden werden die Ergebnisse mehrerer Simulationen dargestellt, die die Effekte möglicher Fehlspezifikationen des propensity score-Modells auf die verschiedenen Adjustierungsmethoden bei der Schätzung des ACE untersuchen sollen.

3.4.1 Simulation 3

Die Kovariablenmatrix, $\mathbf{X}$, enthält die Realisierungen des multivariat normalverteilten Kovariablenvektors $\vec{X}^t = \Big(X_1, \ldots, X_5\Big)$, (vgl. S. 82). Für folgenden Simulationen wird der response-Wert, y_i, durch $y_i = y_{i,c} + \beta_1 \cdot x_{i,1} + \beta_2 \cdot x_{i,4}$, $\beta_1 = 10$, $\beta_2 = 15$ modelliert. Das treatment-assignment der folgenden Simulationen wurde durch das logistische Modell

$$\Pr(S_i = 1 \mid \vec{x}_i) = \frac{1}{1 + \exp(-(\ln(2) \cdot x_{i,1} + \ln(4) \cdot x_{i,4}))}$$

realisiert, so dass mit steigenden X_1-Werten eine zweifache Chance, mit steigenden X_4-Werten eine vierfache Chance einhergeht, zu t selegiert zu werden.Die Korrelationen der Kovariablen mit der response-Variable liegen bei $\rho_{Y_c,X_1} = 0.61$ und $\rho_{Y_c,X_4} = 0.74$, die Korrelation zwischen den beiden konfundierenden Kovariablen bei $\rho_{X_1,X_4} = 0.40$. Abbildung 7 veranschaulicht das zu Grunde gelegte Modell:

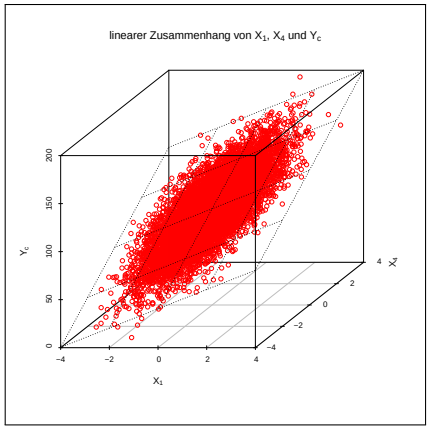

Abbildung 7: Lineares Modell von response-Variable und Kovariablen

In der nachstehenden Simulation wird der lineare Prädiktor des propensity score-Modells in jeder Stichprobe durch

$$\eta_i = \beta_0 + \beta_1 \cdot x_{i,2} + \beta_2 \cdot x_{i,3} + \beta_3 \cdot x_{i,4} + \beta_4 \cdot x_{i,5}$$

geschätzt, so dass bei der Modellierung des linearen Prädiktors die konfundierende Kovariable X_1 nicht berücksichtigt wird (sog. *omitted variables*-Problem).

Tabelle 3: Ergebnisse der Simulation 3

	Methode	*bias*	*Reduk.*	*S.E.*	*M.S.E.*	*Cov.*	*range*
$n = 100$	unadj.	24.36	-	4.58	614.30	0.00	34.84 (6.61, 41.45)
	Dez.-Strat.	5.52	77.34	5.09	56.38	79.38	41.72 (-16.09, 25.63)
	Quint.-Strat.	10.28	57.79	8.11	171.45	73.22	73.13 (-21.76, 51.38)
	IPW_1	6.07	75.08	15.76	285.22	87.82	892.49 (-522.04, 370.46)
	IPW_2	6.71	72.45	5.15	71.55	59.16	68.88 (-41.06, 27.83)
	ex. mat.	6.85	71.88	6.93	94.95	82.82	56.74 (-22.32, 34.42)
$n = 200$	unadj.	24.32	-	3.23	601.90	0.00	22.56 (13.49, 36.05)
	Dez.-Strat.	5.83	76.03	3.36	45.27	58.06	22.70 (-5.41, 17.29)
	Quint.-Strat.	11.96	50.82	6.11	180.37	50.00	49.49 (-11.26, 38.23)
	IPW_1	5.41	77.75	10.44	138.26	80.78	296.26 (-210.32, 85.94)
	IPW_2	6.11	74.88	4.78	60.18	51.06	51.61 (-27.28, 24.32)
	ex. mat.	6.35	73.89	5.01	65.42	75.02	46.76 (-14.25, 32.51)
$n = 500$	unadj.	24.36	-	2.04	597.57	0.00	14.21 (16.78, 30.99)
	Dez.-Strat.	5.99	75.41	2.24	40.90	23.92	16.81 (-1.66, 15.15)
	Quint.-Strat.	13.54	44.42	4.52	203.76	17.62	38.89 (-4.27, 34.62)
	IPW_1	5.11	79.02	8.79	103.38	75.82	175.17 (-113.4, 61.77)
	IPW_2	5.66	76.76	2.74	39.54	37.72	45.85 (-30.26, 15.59)
	ex. mat.	5.87	75.90	3.29	45.28	55.86	25.20 (-5.78, 19.42)

Anm.: ACE = 0, *Reduk.:* prozentuale bias-Reduktion, *Cov.:* prozentuale Coverage, unadj.: unadjustierte Schätzung des ACE, Dez.-Strat.: Dezentil-Stratifikation, Quint.-Strat.: Quintil-Stratifikation, IPW_1 : inverse probability weighting 1, IPW_2 : inverse probability weighting 2, ex. mat.: exaktes matching

Tabelle 3 führt die Ergebnisse der Simulation auf: Sämtliche Adjustierungsverfahren sind grundlegend in der Lage, den vorliegenden bias, der einer unadjustierten Schätzung des ACE folgt, auch dann zu reduzieren, wenn das zu Grunde gelegte propensity score-Modell eine der beiden Störvariablen nicht berücksichtigt; entscheidend ist hierbei, dass die ausgelassene Kovariable X_1 sowie die eingeschlossene Variable X_4 einen linearen Zusammenhang von $\rho_{X_1,X_4} = 0.4$

aufweisen und entsprechend durch die Inklusion von X_4 in das propensity score-Modell Varianzanteile von X_1 vorhanden sind. Auch wenn sämtliche Adjustierungsverfahren bereits in kleinen Stichproben eine erhebliche bias-Reduktion aufweisen, zeigt sich, dass trotz zunehmenden Stichprobenumfangs die resultierende Konfundierung nicht vollständig behoben wird, so dass dem omitted-variables-Problem folgend auch in großen Stichproben ein bias resultiert und die Schätzer kein M.S.E.-konsistentes Verhalten aufweisen. Gegeben einem bias und mit zunehmenden Stichprobenumfang abnehmenden Standardfehlern liegen mit $n = 500$ sämtliche Überdeckungsraten der geschätzten Konfidenzintervalle weit unter dem Vertrauensniveau von 95%: Insbesondere die Quintil-Stratifikation weist hier entsprechende Probleme bei der Adjustierung auf, da mit zunehmenden Stichprobenumfang der resultierende bias, der bereits mit $n = 100$ weit über denen der anderen Adjustierungsmethoden liegt, zunimmt und der zugehörige Standardfehler abnimmt. In Folge dessen resultiert mit $n = 500$ der größte M.S.E. und die geringste coverage im Verhältnis zu den anderen Adjustierungsmethoden.

Ein Vergleich beider weighting-Methoden weist auf, dass insbesondere eine Gewichtung mit dem IPW_2- Schätzer zu Problemen bei Intervallschätzung des ACE führt: In Folge eines abnehmenden Standardfehlers und einer bestehenden Verzerrung nimmt die coverage mit zunehmenden Stichprobenumfang ab, so dass die Überdeckungsrate des IPW_2-Schätzers bei $n = 500$ nur noch bei 37.72% liegt. Die graphische Inspektion der resultierenden Schätzintervalle verdeutlicht das Problem (vgl. Abb. 8). Es ist unmittelbar ersichtlich, dass die Adjustierung mit dem IPW_2-Schätzer zu Intervallschätzungen führt, deren zugehörigen Untergrenzen im extremen negativen Bereich liegen, während derartig extreme Obergrenzen nicht resultieren. Eine solche Tendenz lässt sich für den IPW_1-Schätzer nicht beobachten, da in Folge des allgemein extremen ranges an Punktschätzungen Konfidenzintervalle geschätzt werden, die mit $n = 500$ zu der größten coverage verglichen mit allen anderen Adjustierungsverfahren führt.

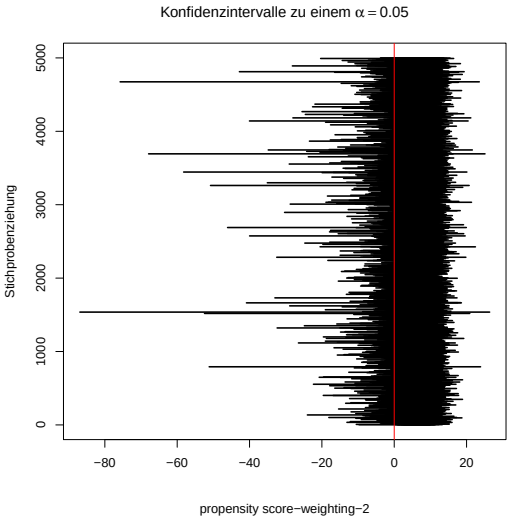

Abbildung 8: 95%−ige Konfidenzintervalle für den IPW_2-Schätzer bei $n = 500$

3.4.2 Simulation 4

Folgende Simulation behält sämtlich obig aufgeführte Spezifikationen bei, jedoch wird folgend der lineare Prädiktor des propensity score-Modells in den Stichproben ohne die konfundierenden Kovariablen X_1 und X_4 geschätzt, so dass: $\eta_i = \beta_0 + \beta_1 \cdot x_{i,2} + \beta_2 \cdot x_{i,3} + \beta_3 \cdot x_{i,5}$.

Ergebnisse

Tabelle 4 führt die Ergebnisse der Simulation auf: Die Konsequenzen bei Ausschluss beider konfundierenden Störvariablen aus dem Modell zur Schätzung des propensity scores sind weitreichend (s. Tabelle 4): keines der Adjustierungsverfahren ist in der Lage, den vorliegenden bias, der einer unadjustierten Schätzung des ACE folgt, zu reduzieren. Sämtliche Adjustierungen resultieren in einem bias, der dem einer unadjustierten Schätzung folgend gleicht. Dem entsprechend liegen sämtliche bias-Reduktionen unterhalb der 10%−Grenze, so dass eine Adjustierung mit den propensity score-Methoden keinen begünstigenden Effekt bei Nivellierung der Konfundierung mit sich bringt. Die Ergebnisse der Quintil-Stratifikation weisen statt einer bias-Reduktion eine Verdoppelung des bias auf, so dass eine Adjustierung mit der Quintil-Stratifikation zu einer größeren Verzerrung führt als im unadjustierten Falle.

In Anbetracht der resultierenden M.S.E.-Schätzungen zeigt sich, dass keines der Adjustierungsverfahren ein M.S.E.-konsistentes Verhalten bei der Schätzung des ACE aufweist: zwar nehmen sämtliche Standardfehler der Adjustierungsmethoden mit zunehmenden Stichprobenumfang ab, die resultierenden Verzerrungen bleiben jedoch unabhängig von der Stichprobengröße bestehen. Für die Quintil-Stratifikation lässt sich eine Zunahme des M.S.E. in Folge des zum zunehmenden n wachsenden bias feststellen. In Folge dessen resultieren für alle Adjustierungsverfahren Intervallschätzungen bei $n = 500$, die mit einer coverage von 0% einhergehen. Hervorzuheben sind die Ergebnisse nach Adjustierung mit dem IPW_1-Schätzer: Während in den vorherigen Simulationen in kleinen Stichproben Punktschätzungen zu beobachten waren, die vorrangig in den extremen negativen Bereich tendierten, folgt einer Adjustierung mit einem propensity score-Modell, das die entscheidenden Störvariablen nicht einschließt, eine Abnahme dieser Tendenz. Die beobachteten Punktschätzungen nach Gewichtung mit dem IPW_1-Schätzer bewegen sich mit $n = 100$ in einem Intervall von $[-17.23, 51.83]$, so dass Punktschätzungen in den extremen positiven Bereich zu beobachten sind, die jedoch nicht im Verhältnis zu vorherigen Simulationen stehen.

Tabelle 4: Ergebnisse der Simulation 4

	Methode	*bias*	*Reduk.*	*S.E.*	*M.S.E.*	*Cov.*	*range*
$n = 100$	unadj.	24.36	-	4.58	614.39	0.00	34.84 (6.61, 41.45)
	Dez.-Strat.	22.18	8.94	5.01	517.05	0.70	37.98 (3.72, 41.69)
	Quint.-Strat.	45.83	-88.13	9.55	2191.59	0.60	91.02 (-0.26, 90.76)
	IPW_1	22.17	8.99	5.68	523.77	4.90	69.05 (-17.23, 51.83)
	IPW_2	22.09	9.32	4.52	508.40	0.48	34.76 (5.83, 40.59)
	ex. mat.	22.22	8.78	6.31	533.54	6.28	46.75 (-2.26, 44.49)
$n = 200$	unadj.	24.32	-	3.23	601.90	0.00	22.56 (13.49, 36.05)
	Dez.-Strat.	22.19	8.76	3.37	503.75	0.00	22.82 (11.38, 34.21)
	Quint.-Strat.	48.83	-100.78	6.91	2432.12	0.00	51.83 (23.26, 74.65)
	IPW_1	22.17	8.84	3.66	504.90	0.20	33.67 (8.97, 42.64)
	IPW_2	22.08	9.21	3.22	497.89	0.00	23.57 (10.56, 34.13)
	ex. mat.	22.17	8.84	4.35	510.43	0.14	33.43 (6.69, 40.12)
$n = 500$	unadj.	24.36	-	2.04	597.57	0.00	14.21 (16.78, 30.99)
	Dez.-Strat.	22.20	8.86	2.08	497.17	0.00	14.04 (15.09, 29.14)
	Quint.-Strat.	48.94	-100.90	4.26	2413.27	0.00	30.89 (32.95, 63.84)
	IPW_1	22.17	8.99	2.21	496.39	0.00	15.59 (14.56, 30.16)
	IPW_2	22.09	9.36	2.05	492.17	0.00	14.43 (14.87, 29.31)
	ex. mat.	22.15	9.07	2.73	498.08	0.00	20.39 (12.43, 32.83)

Anm.: ACE = 0, *Reduk.:* prozentuale bias-Reduktion, *Cov.:* prozentuale Coverage, unadj.: unadjustierte Schätzung des ACE, Dez.-Strat.: Dezentil-Stratifikation, Quint.-Strat.: Quintil-Stratifikation, IPW_1 : inverse probability weighting 1, IPW_2 : inverse probability weighting 2, ex. mat.: exaktes matching

3.4.3 Simulation 5

In der folgenden Simulation wurde der konfundierende Einfluss der Kovariablen X_1 und X_4 durch eine Interaktion beider Variablen modelliert. Der response-Wert wurde mit

$$y_i = y_{i,c} + \beta_1 \cdot (x_{i,1} \cdot x_{i,4}) \quad \beta_1 = 10$$

modelliert, die bedingte Selektionswahrscheinlichkeit durch

$$\Pr(S_i = 1 \mid \vec{x}) = \frac{1}{1 + \exp((-\ln(4) \cdot x_{i,1} \cdot x_{i,4}))}.$$

Abbildung 9 veranschaulicht das zu Grunde gelegte response-Modell. Der lineare Prädiktor des

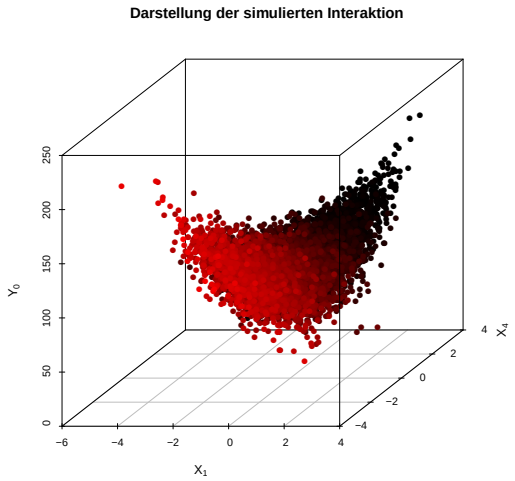

Abbildung 9: Modell von response-Variable und Kovariablen

propensity score-Modells wurde ohne Rücksicht auf die Interaktion modelliert, so dass

$$\eta_i = \beta_0 + \beta_1 \cdot x_{i,1} + \beta_2 \cdot x_{i,2} + \beta_3 \cdot x_{i,3} + \beta_4 \cdot x_{i,4} + \beta_5 \cdot x_{i,5}.$$

Ergebnisse

Tabelle 5: Ergebnisse der Simulation 5

	Methode	*bias*	*Reduk.*	*S.E.*	*M.S.E.*	*Cov.*	*range*
$n = 100$	unadj.	8.17	-	3.67	80.22	-	25.96 (-4.61, 21.11)
	Dez.-Strat.	7.73	5.39	4.02	75.91	52.34	29.43 (-6.97, 22.45)
	Quint.-Strat.	16.39	-100.61	7.71	328.08	44.22	62.52 (-15.62, 46.90)
	IPW_1	8.28	-1.35	5.04	93.96	59.34	50.89 (-24.71, 26.18)
	IPW_2	8.13	0.49	3.51	78.42	37.02	24.96 (-4.30, 20.66)
	ex. mat.	7.92	3.06	4.85	86.25	62.88	35.56 (-8.95, 26.61)
$n = 200$	unadj.	8.23	-	2.59	74.44	-	18.77 (-0.76, 18.01)
	Dez.-Strat.	7.98	3.04	2.72	71.08	15.42	20.36 (-1.73, 18.62)
	Quint.-Strat.	18.09	-119.81	5.61	358.72	11.24	44.82 (-1.77, 43.05)
	IPW_1	8.31	-0.97	2.92	77.58	16.18	21.11 (-1.95, 19.16)
	IPW_2	8.23	0.00	2.51	74.03	9.18	19.96 (-1.22, 18.74)
	ex. mat.	8.10	1.58	3.30	76.50	33.04	21.78 (-2.11, 19.67)
$n = 500$	unadj.	8.26	-	1.64	70.92	0.00	11.64 (1.94, 13.58)
	Dez.-Strat.	8.03	2.78	1.67	67.27	0.30	12.51 (1.02, 13.53)
	Quint.-Strat.	18.20	-120.34	3.44	343.07	0.08	26.71 (4.28, 30.99)
	IPW_1	8.29	-0.36	1.67	71.51	0.12	11.58 (2.39, 13.97)
	IPW_2	8.26	0.00	1.61	70.82	0.06	11.33 (2.14, 13.47)
	ex. mat.	8.07	2.30	2.07	69.41	3.20	15.32 (0.62, 15.95)

Anm.: ACE = 0, *Reduk.:* prozentuale bias-Reduktion, *Cov.:* prozentuale Coverage, unadj.: unadjustierte Schätzung des ACE, Dez.-Strat.: Dezentil-Stratifikation, Quint.-Strat.: Quintil-Stratifikation, IPW_1 : inverse probability weighting 1, IPW_2 : inverse probability weighting 2, ex. mat.: exaktes matching

Keines der Adjustierungsverfahren ist in der Lage, den bias, der einer unadjustierten Schätzung des ACE folgt, im auffälligen Ausmaße zu reduzieren. Für $n = 200$ und $n = 500$ entspricht der geschätzte bias des IPW_2-Schätzers dem der unadjustierten Schätzung, so dass eine Adjustierung an dieser Stelle keinen begünstigenden Effekt aufweist. Sowohl die Quintil-Stratifikation

als auch die Gewichtung mit der IPW_1-Methode führen zu Schätzungen des ACE, die eine größere Verzerrung aufweisen als im unadjustierten Falle. Damit ist ersichtlich, dass eine Fehlspezifikation des lineares Prädiktors, der zur Schätzung der propensity scores benutzt wird, zu Ergebnissen führt, die der ursprünglichen Motivation des propensity scores gegenüber stehen.

Einerseits nehmen die Standardfehler aller Adjustierungsmethoden mit zunehmenden Stichprobenumfang ab, andererseits bleiben die resultierenden Verzerrungen bestehen, so dass in Folge die Überdeckungsraten bei $n = 500$ fast vollständig gegen null tendieren. Zusätzlich resultiert hieraus, dass keines der Adjustierungsverfahren ein konsistentes Verhalten ausweist: Zwar lässt sich festhalten, dass die geschätzten M.S.E.-Werte mit zunehmenden zum n geringfügig abnehmen, jedoch lässt sich diese Tendenz vollständig durch die abnehmenden Standardfehler erklären, da entsprechende Verzerrungen weiterhin zu beobachten sind.

3.5 Diskussion der Ergebnisse

Die ersten beiden Simulationen weisen eindeutige Ergebnisse auf: In diesen Simulationen wurde der propensity score in den Stichproben derartig geschätzt, dass die jeweilige konfundierende Störvariable zusammen mit allen anderen erhobenen Kovariablen in das logistische Regressionsmodell aufgenommen wurde. Die adjustierten Schätzungen zeigen auf, dass die Konfundierung bei der Schätzung des ACE größtenteils behoben werden kann, auch wenn bei Vorliegen einer stetigen Störvariablen eine Vielzahl der Adjustierungsmethoden erst mit zunehmenden n sich dem ACE annäherende Ergebnisse aufweisen.

Bei Vorliegen einer diskreten Störvariablen (Simulation 1) lässt sich bereits in kleinen Stichproben empirisch eine deutliche bias-Reduktion festhalten: sämtliche Adjustierungsverfahren weisen mit $n = 100$ eine bias-Reduktion von mindestens 93% auf und es resultieren Punktschätzungen, die im Durchschnitt approximativ mit dem ACE übereinstimmen. In Anbetracht der abnehmenden Standardfehler und M.S.E.-Werte bei zunehmenden Stichprobenumfang lässt sich zusätzlich ein M.S.E.-konsistentes Verhalten festhalten. Die einzige Ausnahme hier stellt die Quintil-Stratifikation dar: Ausgehend von den theoretischen Überlegungen, dass mit entsprechend grober Stratifikation die propensity scores innerhalb eines stratums stärker streuen und folglich eine Konfundierung der Störvariablen bestehen bleibt, handelt es sich um dasjenige Adjustierungsverfahren mit dem größten M.S.E. und der geringsten coverage bei $n = 500$. Der Theorie entsprechend lassen sich in kleinen Stichproben für das exakte matching

sowie die IPW_1-Methode verhältnismäßig große Standardfehler festhalten, die auf einen entsprechend gewichtigen Einfluss extremer Gewichte bzw. geringe matched paires zurückzuführen sind; mit $n = 500$ lassen sich Standardfehler festhalten, die denen der anderen Adjustierungsverfahren entsprechen.

Bei Vorliegen einer stetigen Störvariablen (Simulation 2) resultieren in kleinen Stichproben bias-Reduktionen von mindestens 83%, die eine entsprechende Nutzung der propensity score-Methoden weiterhin rechtfertigen, wobei die bestehenden Verzerrungen nach Adjustierung im Verhältnis zur ersten Simulation auffällig groß sind. Bis auf die Dezentil-Stratifikation decken sich mit $n = 100$ die durchschnittlichen Punktschätzungen nicht mit dem ACE, auffällig sind hier insbesondere die Verzerrungen des exakten matchings und des IPW_2-Schätzers. Für das exakte matching lässt sich die resultierende Verzerrung in kleinen Stichproben eindeutig auf die normalverteilte Störvariable zurückführen, die ein exaktes matching in Folge der stetigen Realisierungen erschwert. Es zeigt sich für das exakte matching zum Einen, dass trotz der bestehenden Verzerrung die Überdeckungsraten der Intervallschätzungen mit der festgelegten coverage vereinbar sind, zum Anderen, dass der bias mit zunehmenden Stichprobenumfang abnimmt, da die entsprechende Wahrscheinlichkeit für einen match zunimmt. Explizit der IPW_1-Schätzer erweist sich in großen Stichproben hinsichtlich der Verzerrung allen anderen Adjustierungsverfahren überlegen, da die durchschnittlichen Punktschätzungen mit $n = 500$ approximativ dem ACE gleichen; in Anbetracht des großen Standardfehlers handelt es sich jedoch um das Verfahren, das auch in großen Stichproben mit einer entsprechenden großen Unsicherheit bei der Schätzung des ACE einhergeht.

Diejenigen Simulationen, in denen eine Fehlspezifikation des propensity score-Modells zu Grunde gelegt wird, weisen die „paradoxical nature of the propensity score“ (Imai & Ratkovic, 2014, S.244) auf: Die Verwendung des propensity scores führt bei richtiger Spezifikation des Schätzmodells zu einer Nivellierung der Konfundierung der Störvariablen, bei Fehlspezifikation jedoch zu inkonsistenten (Simulation 3) Schätzergebnissen und resultierenden bias-Werten, die z.T. größer sind als die der unkonfundierten Schätzung (Simulation 4 & 5).

Das in der Simulation 3 spezifizierte propensity score-Modell beinhaltete eine der beiden Störvariablen und zwischen beiden Kovariablen lag ein moderater linearer Zusammenhang von $\rho_{X_1,X_4} = 0.4$ vor, so dass durch die in das Modell aufgenommene Variable entsprechend Varianzanteile der ausgeschlossenen Variable enthalten waren. Dementsprechend ließen sich bereits

in kleinen Stichproben bias-Reduktionen von ca. 70% beobachten; einzige Ausnahme stellt an dieser Stelle die Quintil-Stratifikation dar. Im Verhältnis zu den vorherigen Simulationen weisen die Schätzer jedoch kein M.S.E.-konsistentes Verhalten auf: zwar nehmen die Standardfehler mit zunehmenden Stichprobenumfang ab, jedoch liegen die mit $n = 100$ resultierenden Verzerrungen weiterhin mit $n = 500$ vor. Diesem Verhalten entsprechend nehmen mit zunehmenden Stichprobenumfang die Überdeckungsraten ab und für den IPW_1-Schätzer mit der größten coverage lässt sich bei $n = 500$ eine Überdeckungsrate von nur noch ca. 76% festhalten.

Bei vollständiger Auslassung der beiden Kovariablen (Simulation 4) resultieren Verzerrungen, die sich nicht auffällig von der Verzerrung bei unadjustierter Schätzung unterscheiden; für die Quintil-Stratifikation lässt sich stattdessen ein größerer bias als der bei unadjustierter Schätzung des ACE festhalten. Selbige Ergebnisse lassen sich bei fehlender Modellierung der Interaktion aufweisen: Damit ist eindeutig, dass eine Adjustierung mit dem propensity score nur zu einer auffälligen bias-Reduktion und konsistenten Schätzung des ACE führt, wenn das entsprechende Schätzmodell des propensity scores die konfundierenden Störvariablen inkludiert.

4 Non-parametrische Schätzung des Propensity Scores

4.1 Einleitung

De facto ist der propensity score, $e(\vec{x}_i)$, in einer observational study ein unbekannter Skalar und muss ausgehend von den beobachteten Daten $(s_i, \vec{x}_i)$ geschätzt werden. Die Relevanz einer korrekten Spezifikation des treatment-assignments bei parametrischer Schätzung des propensity scores über das logistische Regressionsmodell wurde in den vorherigen Simulationen deutlich: Sobald das zur Schätzung des propensity scores verwendete Modell nicht mehr vollständig und korrekt spezifiziert ist, weisen die Adjustierungsmethoden, die auf dem geschätzten propensity score basieren, bei der Schätzung des ACE eine auffällige Verzerrung sowie ein M.S.E.-inkonsistentes Verhalten auf.

Als Alternative zur parametrischen Modellierung schlugen McCaffrey et al. (2004) ein non-parametrisches Schätzverfahren für den propensity score vor, das auf der Idee der *boosted logistic regression* und des *regression trees* basiert und keine Modellierung des treatment-assignments benötigt. Stattdessen wird der propensity score in einem iterativ spezifizierten logistischen Regressionsmodell dergestalt geschätzt, dass die in das finale Modell aufgenommenen, den regression trees resultierenden Kovariablen diejenigen sind, die die Bernoulli log-likelihood maximieren. Jedoch besteht trotz eines maximalen Modell-Fits, hier indiziert durch die Bernoulli log-likelihood, stets die Möglichkeit einer Fehlspezifikation des Modell und in Konsequenz würden die Adjustierungsmethoden den ACE verzerrt und inkonsistent schätzen (Imai & Ratkovic, 2014).

Im Folgenden wird ein non-parametrischer Schätzer für den propensity score vorgeschlagen, der sich für den speziellen Fall, dass die Realisierungen der Kovariablen in $\mathbf{X}$ stetig und die zugehörigen Dichten log-konkav sind, anwenden lässt. Für die Definition des Schätzers wird im ersten Schritt der propensity score über den Satz von Bayes umgeformt, um die Identifikation des propensity scores durch die unbekannten (bedingte) Dichten aufzeigen zu können, im zweiten Schritt wird ein non-parametrischer Schätzer für die Klasse der log-konkaven Dichten, der von Cule et al. (2010) vorgeschlagen wurde, vorgestellt, der zur Schätzung der unbekannten bedingten Dichten im Folgenden genutzt wird. Anhand mehrerer Simulationen soll gezeigt werden, dass eine Adjustierung mit dem geschätzten propensity scores zu einer bias-Reduktion bei der Schätzung des ACE führt und es soll geprüft werden, ob die resultierenden Schätzungen

konsistent sind.

4.1.1 Propensity Score und der Satz von Bayes

Wie bereits in (21) aufgewiesen, lässt sich der propensity score der Einheit i, definiert als die bedingte Wahrscheinlichkeit, gegeben dem Kovariablenvektor $\vec{x}_i$ zum treatment selegiert zu werden, über den Satz von Bayes darstellen. Es gilt mit $s_i \in \{0, 1\}$:

$$e(\vec{x}_i) := \Pr(S_i = 1 \mid \vec{x}_i) = \frac{\Pr(\vec{x}_i \mid S_i = 1) \cdot \Pr(S_i = 1)}{\Pr(\vec{x}_i \mid S_i = 1) \cdot \Pr(S_i = 1) + \Pr(\vec{x}_i \mid S_i = 0) \cdot \Pr(S_i = 0)},$$

im Falle stetiger Kovariablen entsprechend

$$e(\vec{x}_i) := \frac{f(\vec{x}_i \mid S_i = 1) \cdot \Pr(S_i = 1)}{f(\vec{x}_i \mid S_i = 1) \cdot \Pr(S_i = 1) + f(\vec{x}_i \mid S_i = 0) \cdot \Pr(S_i = 0)}, \tag{47}$$

wobei $f(\vec{x}_i \mid S_i = s_i)$ die an $S_i = s_i$ bedingte, ggf. multivariate, und in praxi zumeist unbekannte Dichtefunktion der Kovariablen indiziert.

Gegeben der Umformung über den Satz von Bayes ist eine mögliche Schätzung des propensity scores in Folge von Schätzungen der Selektionswahrscheinlichkeiten, $\Pr(S_i = 1)$ und $\Pr(S_i = 0)$, sowie der bedingten Dichten, $f(\vec{x}_i \mid S_i = s_i)$, ersichtlich. In großen Stichproben lässt sich die unbekannte Wahrscheinlichkeit $\Pr(S_i = 1)$ konsistent schätzen über den Anteil der Einheiten, die t zugeordnet ist, d.h.: $\hat{\pi}_i = \frac{n_t}{n}$, vice versa für $\Pr(S_i = 0)$. Einer konsistenten Schätzung der bedingten Dichten, $f(\vec{x}_i \mid S_i = s_i)$, folgend wäre der propensity score von i vollständig identifiziert und konsistent schätzbar.

Es liegen verschiedene Schätzer für eine unbekannte, ggf. multivariate, Dichte vor, wobei insbesondere die Klasse der non-parametrischen *Kerndichte-Schätzer* von Interesse ist, da diese eine gleichmäßig stetige und konsistente Schätzung der Dichtefunktion ermöglichen (z.B., Epanechnikov, 1969; Parzen, 1962), solange die *Bandbreite* des Kerndichte-Schätzers optimal gewählt ist. Damit liegt ein entscheidender Nachteil bei Verwendung einer der möglichen Kerndichte-Schätzer vor: Die Approximation eines Kerndichte-Schätzers an die Dichte ist stets abhängig von der optimalen Wahl der Bandbreite, auch wenn diesbezüglich Algorithmen zur Optimierung vorgeschlagen werden (vgl. hierzu Cule et al., 2010). Zusätzlich unterliegt die Schätzung einer Dichtefunktion durch die Kerndichte-Schätzer dem Phänomen der *curse of dimensionality*, so dass insbesondere in kleinen Stichproben bei der Schätzung höher-dimensionaler Dichtefunktionen Probleme bei der Approximation auftreten können.

Cule et al. (2010) beweisen in ihrem Artikel die Existenz eines eindeutigen Maximum Likelihood-Schätzers, $\hat{f}_n$, für die Klasse der (multivariaten) log-konkaven Dichten, der diese auch ohne Spezifikation einer Bandbreite konsistent schätzt. Der Vorteil dieses Dichteschätzers soll folgend hervorgehoben werden und in einen Zusammenhang mit der Schätzung des propensity scores gebracht werden.

4.1.2 Exkurs: Log-konkaver Dichteschätzer

Ausgehend von dem bestehenden Problem, dass der Verwendung eines Kerndichte-Schätzers stets die Wahl einer entsprechenden Bandbreite vorangeht und die Güte der Schätzung abhängig von der Wahl der Bandbreite ist, schlugen Cule et al. (2010) einen Maximum Likelihood-Schätzer vor, der ohne die Spezifikation eines Bandbreite-Parameters auskommt, und eine unbekannte, ggf. multivariate Dichte f automatisch, non-parametrisch und konsistent schätzt, solange die Dichtefunktion f der Klasse der log-konkaven Dichten zugehört. Somit wird folgend vorausgesetzt, dass $\log(f)$ eine konkave Funktion ist; allgemein ist eine Funktion f, $f : D \to \mathbb{R}$ mit einer konvexen Menge $D \subset \mathbb{R}^n$ und $f(x) > 0, \forall x \in D$, log-konkav, wenn $\forall x, y \in D$ und $\forall \lambda \in [0, 1]$ gilt:

$$\log(f(\lambda x + (1 - \lambda)y)) \geq \lambda \log(f(x)) + (1 - \lambda) \log(f(y)).$$

Eine alternative Formulierung kann wie folgt geschehen: Eine Funktion f, $f : D \to \mathbb{R}$, heißt konkav, wenn ihr Hypograph,

$$\text{hyp } f := \{(x, \mu) \mid f(x) \geq \mu\},$$

eine konvexe Menge ist.

Insbesondere in den Human- und Sozialwissenschaften scheint die Verwendung des log-konkaven Dichteschätzers naheliegend, da eine Vielzahl zugehöriger Variablen und Konstrukte mit Dichten modelliert werden, die log-konkav sind; Beispiele hierfür sind die Normalverteilung, die Exponentialverteilung, die logistische Verteilung, die χ^2-Verteilung (mit $df \geq 2$) oder die Student's t-Verteilung (mit Ausnahme bestimmter Freiheitsgrade). Abbildung 10 veranschaulicht beispielhaft das Prinzip der log-Konkavität für die Dichte einer bivariate Standardnormalverteilung mit $\rho = 0.8$; zusätzlich zur Dichte abgetragen ist der natürliche Logarithmus der Dichte. Es ist unmittelbar ersichtlich, dass die logarithmierte Dichte konkav ist, da der zugehörige Graph oberhalb jeder Verbindungsstrecke zweier möglicher Punkte verläuft.

Mathematisch lässt sich dies für den univariaten Fall durch die zweite Ableitung der logarithmierten Dichte der Standardnormalverteilung nach x zeigen, da diese negativ ist:

$$
\begin{aligned}
f(x) &= (2\pi)^{-\frac{1}{2}} \exp\left(-\frac{x^2}{2}\right) \\
\log\left(f(x)\right) &= -\frac{1}{2}\log(2\pi) - \frac{1}{2}x^2 \\
\log\left(f(x)\right)' &= -x \\
\log\left(f(x)\right)'' &= -1
\end{aligned}
$$

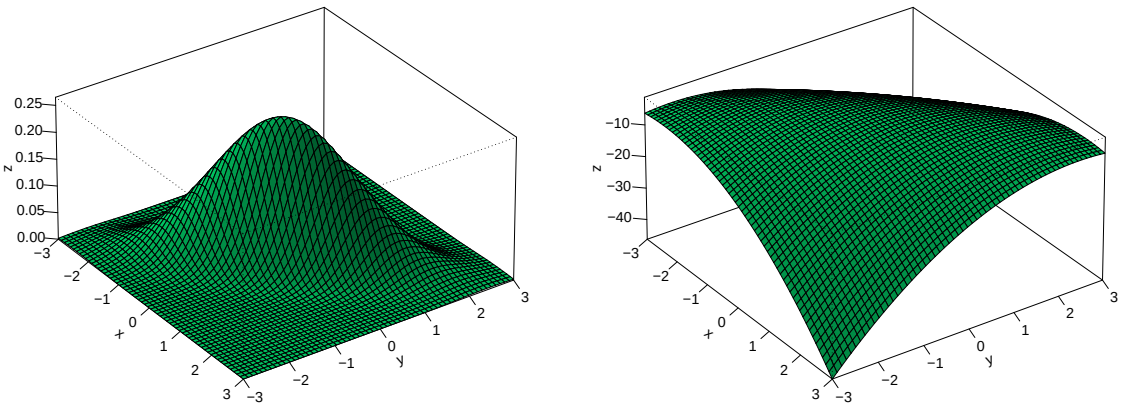

Abbildung 10: Dichte und Logarithmus der Dichte einer bivariaten Standardnormalverteilung

Cule et al. (2010) beweisen in ihrem Artikel die Existenz eines eindeutigen ML-Schätzers für log-konkave Dichten, $\hat{f}_n$, ausgehend von der Überlegung, dass log-konkave Dichten stets unimodal sind. Folglich stellen im übertragenen Sinne alle Werte der logarithmierten Dichte ein „Zelt“ dar. Jedem Dichtewert als Punkt auf diesem Zelt lässt sich eine Höhe, folgend $\vec{y}^t = (y_1, \ldots, y_n)$, zuordnen, die den Abstand dieses Wertes zu den $\vec{x}-$Werten angibt und mit einer zugehörigen „tent function“ (Cule et al., 2010, S.2), $\bar{h}_y$, beschrieben werden kann, die die Existenz des ML-Schätzers entsprechend begründet, wobei $\bar{h}_y$ die kleinste konkave Funktion ist, die die Punkte $(X_1, y_1), \ldots, (X_n, y_n)$ in ihrem Hypographen hat. Der ML-Schätzer ist definiert als Stelle des Maximums der Funktion $\psi(f)$:

$$\psi_n(f) = \frac{1}{n}\sum_{i=1}^{n} \log(f(X_i)) - \int_{\mathbb{R}^n} f(x)dx.$$

Der ML-Schätzer wird mit Hilfe folgender Funktion bestimmt:

$$\sigma(y_1, \ldots, y_n) = -\frac{1}{n}\sum_{i=1}^{n} y_i + \int_{C_n} \exp\{\bar{h}_y(x)\}dx,$$

wobei C_n die *konvexe Hülle* der Variablen $X_1, \ldots, X_n$, d.h. $C_n = \text{conv}(X_1, \ldots, X_n)$, beschreibt. Für die Funktion $\sigma(y_1, \ldots, y_n)$ wird y^*, d.h. die Stelle an der $\sigma(y_1, \ldots, y_n)$ minimal wird, bestimmt. Damit gilt (Cule et al., 2010, S.9): $\log(\hat{f}_n) := \bar{h}_{y^*}$, mit anderen Worten ist $\log(\hat{f})$ identisch mit einer tent-Funktion.

In einer Reihe von Simulationsstudien zeigen die Autoren, dass der vorgeschlagene ML-Schätzer unterschiedliche Dichten, z.B. die einer univariaten Normalverteilung, einer bivariaten Normalverteilung mit oder ohne linearen Zusammenhang, einer Gamma-Verteilung, etc., stets mit einem geringeren M.S.E. als der Gauß-Kerndichte-Schätzer mit optimaler Spezifikation der Bandbreite schätzt und ein konsistentes Verhalten zeigt.

Der Zusammenhang zur Schätzung des unbekannten propensity scores, $e(\vec{x}_i)$, lässt sich unmittelbar herstellen: Gegeben der Annahme, $\vec{x}_i$ enthält ausschließlich die Realisierungen stetiger Zufallsvariablen mit unbekannter, multivariater Dichte $f(\vec{x})$, so lässt sich der propensity score nach dem Satz von Bayes durch (47) darstellen; damit sichtlich ist, dass der propensity score von diesen unbekannten, bedingten Dichten und den Selektionswahrscheinlichkeiten abhängt. Da

$$\frac{n_t}{n} \xrightarrow{f.s.} \Pr(S_i = 1),$$
$$\frac{n_c}{n} \xrightarrow{f.s.} \Pr(S_i = 0)$$

und

$$\hat{f}_n \xrightarrow{f.s.} f(\vec{x}),$$

sollte einer konsistenten Schätzung der Selektionswahrscheinlichkeiten, $\Pr(S_i = 1)$ und $\Pr(S_i = 0)$, sowie der (multivariaten) bedingten Dichte $f(\vec{x}_i \mid S_i = s_i)$ folgend der propensity score entsprechend konsistent geschätzt werden können. Folglich sollte unter Anwendung einer der aufgewiesenen Adjustierungsmethoden, die den propensity score nutzen, eine Reduktion des bias, der einer unadjustierten Schätzung des ACE folgt, möglich sein und zusätzlich sollten diese Schätzungen mit zunehmenden Stichprobenumfang gegen den ACE konvergieren. Dies gilt es folgend in vier Simulationen zu zeigen, wobei jeweilig verschiedene Verteilungsmodelle der konfundierenden Störvariablen zu Grunde gelegt wird.

4.2 Simulationen

4.2.1 Methodik und Aufbau der Simulationen

In jeder der folgenden vier Simulationen wurde eine finite Grundgesamtheit mit $N = 10.000$ Einheiten definiert; jede statistische Einheit besitzt zwei potentielle response-Werte, $y_{i,t}$ und $y_{i,c}$, die den response unterhalb der Experimentalbedingung, zu der i nach Stichprobenziehung selegiert wird, angeben. Als Verteilungsmodell für die potentiellen response-Werte wurde $Y_c \sim \mathcal{N}(100, 15^2)$ sowie $Y_t = Y_c$ gewählt, damit: ACE $= E(Y_t) - E(Y_c) = 0$. Zusätzlich wird jeder Einheit i ein Kovariablenvektor $\vec{x}_i$ zugewiesen, der folgend für die Szenarien konkretisiert wird.

Der beobachtete response der Einheit i wird folgend durch ein lineares Modell, das zusätzlich einen linearen Einfluss der Kovariablenwerte auf den potentiellen response-Wert modelliert, spezifiziert; mit ACE $= 0$ gilt allgemein für die Simulationen: $y_i = y_{i,c} + \vec{x}_i^t \vec{\beta}$.

In den vier Simulationen wurden folgende Modelle für den response-Wert sowie für das treatment-assignment zu Grunde gelegt:

In der ersten Simulation soll eine adjustierte Schätzung des ACE mit der vorgeschlagenen propensity score-Schätzung untersucht werden, wobei in dieser Simulation die Dichte der Störvariable univariat und log-konkav ist. $\vec{x}_i$ enthält als Eintrag die Realisierung einer standardnormalverteilten Zufallsvariablen X_1, es gilt: $X_1 \sim \mathcal{N}(0, 1^2)$ und $\beta_1 = 10$. Das treatment-assignment, das nach Stichprobenziehung einsetzt, wurde durch ein logistisches Modell realisiert mit $\Pr(S_i = 1 \mid \vec{x}_i) = F(\log(1.5) \cdot x_{1,i})$, wobei $F(\cdot)$ die Verteilungsfunktion einer logistischen Zufallsvariablen beschreibt.

In der zweiten Simulation sollte das Verhalten des log-konkaven Dichteschätzers, bzw. des resultierenden propensity scores, bei multivariater und log-konkaver Dichte der konfundierenden Störvariablen untersucht werden. Hierzu wurde zunächst eine multivariate Normalverteilung, $\vec{X} \sim \mathcal{N}(\vec{0}, \mathbf{\Sigma})$, mit der positiv-definiten Kovarianzmatrix

$$\mathbf{\Sigma} := \begin{pmatrix} 1 & 0.5 & 0.3 & -0.2 \\ 0.5 & 1 & 0.1 & 0.4 \\ 0.3 & 0.1 & 1 & 0.2 \\ -0.2 & 0.4 & 0.2 & 1 \end{pmatrix}$$

definiert, $\vec{x}_i$ enthielt bei der Modellierung des beobachteten response-Wertes y_i die Realisierung der Variablen X_1 und es wurde $\beta_1 = 10$ gesetzt. Das treatment-assignment wurde in diesem

Szenario durch $\Pr(S_i = 1 \mid \vec{x}_i) = F(\log(2) \cdot x_{1,i} + \log(3) \cdot x_{2,i} + \log(3) \cdot (x_{3,i} \cdot x_{4,i}))$ realisiert; damit unterliegt die Selektion einer Interaktion der Kovariablen X_3 und X_4.

In der dritten Simulation enthielt $\vec{x}_i$ als Eintrag die Realisierung einer univariaten, *heavy-tailed* Student's $t-$Verteilung, $X_1 \sim T(3)$; es gilt $\beta_1 = 10$. Das treatment-assignment wurde durch ein logistisches Modell realisiert mit $\Pr(S_i = 1 \mid \vec{x}_i) = F(\log(1.5) \cdot x_{1,i})$.

Um zu prüfen, wie robust der log-konkave Dichteschätzer auf eine Verletzung der log-Konkavität bei der Schätzung der Dichtefunktion reagiert, enthielt der Kovariablenvektor $\vec{x}_i$ die Realisierung einer Pareto-verteilten Zufallsvariablen, $X_1 \sim \text{Par}(5, 5)$, deren Dichte nicht log-konkav ist; es gilt $\beta_1 = 10$. Allgemein besitzt eine Pareto-verteilte Zufallsvariable, $\text{Par}(k, x_{min})$, mit den Parametern $k > 0$ und $x_{min} > 0$, die Dichte

$$f(x) = \begin{cases} \frac{k \cdot x_{min}^k}{x^{k+1}} & \text{für } x \geq x_{min} \\ 0 & \text{für } x < x_{min}. \end{cases}$$

Die Dichtefunktion ist nicht log-konkav, wie die zweite Ableitung der logarithmierten Dichte nach x zeigt:

$$\begin{aligned} \log(f(x)) &= \log\left(\frac{k \cdot x_{min}^k}{x^{k+1}}\right) \\ &= \log\left(k \cdot x_{min}^k\right) - \log\left(x^{k+1}\right) \\ &= \log(k) + k \cdot \log\left(x_{min}\right) - (k+1) \cdot \log(x) \\ \log(f(x))' &= -\frac{(k+1)}{x} \\ \log(f(x))'' &= \frac{(k+1)}{x^2} \end{aligned}$$

Abbildung 11 stellt die Dichte sowie die logarithmierte Dichte der verwandten Pareto-Variablen, $\text{Par}(5, 5)$, für das Intervall $X \in [0, 20]$ dar und verdeutlicht die Verletzung der Annahme.

Jede der vier Simulationen wurde mit den Stichprobenumfängen $n \in \{250, 500, 1000\}$ jeweilig $k = 400$ Mal wiederholt. Der einfachen Stichprobenziehung der n Einheiten folgend setzte das treatment-assignment ein - die jeweiligen Selektionsmodelle wurden oben spezifiziert. Nach Selektion der gezogenen Einheiten in eine der Bedingungen wurde zur empirischen Ermittlung des bias, der ohne Adjustierung der Kovariablen resultiert, in jeder Stichprobe als Schätzer für den ACE $\bar{Y}_t - \bar{Y}_c$ ermittelt; das resultierende Mittel dieser Punktschätzungen dient als Evidenz für den bias bei unadjustierter Schätzung.

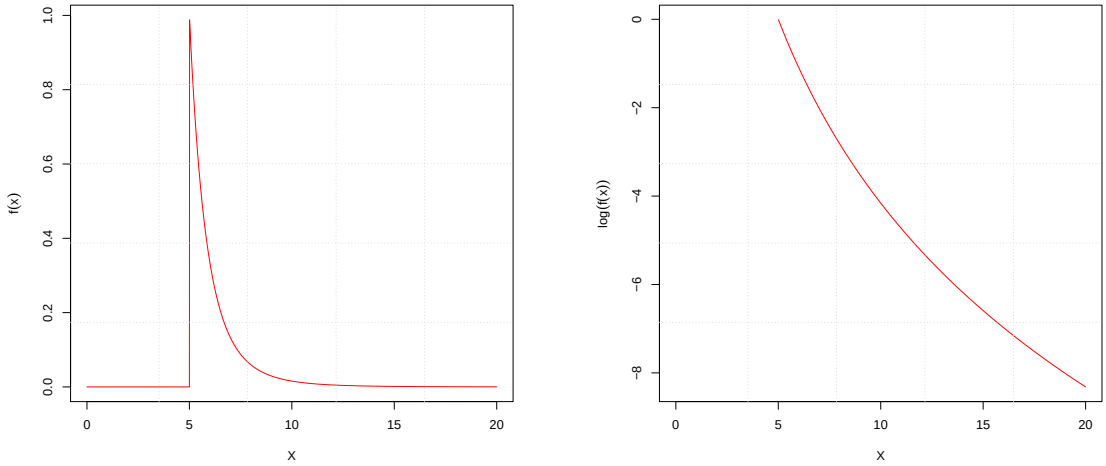

Abbildung 11: Dichte und Logarithmus der Dichte für Par$(5, 5)$

Der unadjustierten Schätzung des ACE folgend wurden zwei propensity scores in jeder Stichprobe geschätzt: zum Einen wurde die vorgeschlagene non-parametrische Schätzung des propensity scores vorgenommen, indem zunächst unterhalb beider Experimentalbedingungen, t und c, die bedingten Dichten der Kovariablen $\vec{x}_i$ über den log-konkaven Dichteschätzer sowie die Selektionswahrscheinlichkeiten über die resultierenden Stichprobenanteile geschätzt worden sind. Zur Ermittlung des propensity scores, folgend $\hat{e}(\vec{x}_i)_{\log}$, wurden die geschätzten Dichtewerte und Selektionswahrscheinlichkeiten in (47) eingesetzt. Zum Anderen wurde zusätzlich in jeder Stichprobe in einem logistischen Regressionsmodell das zu Grunde liegende treatment-assignment modelliert, um zu prüfen, ob der non-parametrisch geschätzte propensity score gegen den tatsächlichen konvergiert. Mit beiden propensity scores wurde der Einfachheit halber eine Quintil-Stratifikation durchgeführt; in jedem stratum wurde die stratumsspezifische Mittelwertsdifferenz ermittelt und der ACE durch das gewogene Mittel der stratumsspezifischen Schätzungen geschätzt. Die Varianzschätzung des ACE erfolgt identisch zu der ersten Simulationsstudie.

Als Gütemaße der Schätzungen werden für den unadjustierten sowie für die beiden adjustierten Fälle für alle n gesondert aufgeführt:

a) der resultierende, empirische bias; indiziert durch das Mittel der jeweiligen Punktschätzungen, $(m(\hat{\tau}))$,

b) der geschätzte M.S.E. mit M.S.E. $= \text{bias}(\hat{\tau})^2 + \text{Var}(\hat{\tau})$,

c) die resultierende, prozentuale bias-Reduktion, definiert als

$$\frac{\text{bias}_{\text{adjustiert}} - \text{bias}_{\text{unadjustiert}}}{\text{bias}_{\text{unadjustiert}}},$$

d) die prozentuale coverage als der prozentuale Anteil an 95%−igen Stichproben-Konfidenzintervallen, die den ACE überdecken,

e) Minimum und Maximum der Punktschätzungen.

Zusätzlich werden die gemittelten, absoluten Differenzen zwischen dem tatsächlichen und non-parametrisch geschätzten propensity score für jedes n angegeben, um eine mögliche Konvergenz empirisch aufzuweisen. Die Ergebnisse der Simulationen werden in Tabellen sowie, wenn aussagekräftig, durch Konfidenzintervalle graphisch präsentiert.

4.2.2 Univariate Normalverteilung

Die in Tabelle 6 abgetragenen Ergebnisse verdeutlichen, dass bereits in kleinen Stichproben einer Adjustierung mit dem non-parametrisch geschätzten propensity score eine Reduktion von fast 96% des bias, der auf die konfundierende Störvariable zurückzuführen ist, bei Schätzung des ACE folgt. Damit folgt einer Adjustierung mit dem non-parametrisch geschätzten propensity score eine größere bias-Reduktion als mit dem parametrisch und korrekt spezifizierten propensity score; ein Befund, der sich auf alle Stichprobenumfänge verallgemeinern lässt.

Sowohl für die Adjustierungen mit dem parametrisch als auch mit dem non-parametrisch geschätzten propensity score lässt sich festhalten, dass der bias bei Schätzung des ACE mit zunehmenden Stichprobenumfang geringfügig zunimmt; ein Umstand, der auf die Adjustierungsmethode der Stratifikation zurückgeführt werden kann, da mit zunehmenden n größere Varianzen des propensity scores innerhalb eines stratums und damit bestehen bleibende Konfundierungen durch die Störvariable einhergehen.

Sowohl die resultierenden Minima und Maxima der Punktschätzungen als auch die M.S.E.-Schätzungen weisen auf, dass mit zunehmenden n die zugehörigen Schwankungen bei der Schätzung des ACE nach Adjustierung mit dem non-parametrisch geschätzten propensity score abnehmen. Für alle n sind die Minima und Maxima der Punktschätzungen nach Adjustierung mit $\hat{e}(\vec{x}_i)_{\log}$ vergleichbar mit denen nach Adjustierung mit dem theoretischen propensity score.

Tabelle 6: Ergebnisse der Simulation 1

		unadjustiert	$\hat{e}(\vec{x}_i)_{\log}$	$e(\vec{x}_i)$
$n = 250$	bias	3.84	0.17	0.33
	[min., max.]	[-2.36, 10.48]	[-6.10, 6.89]	[-5.42, 6.70]
	red.		95.57%	91.41%
	M.S.E.	20.37	4.31	4.08
	cov.	58.75%	96.25%	95.00%
$n = 500$	bias	3.83	0.27	0.40
	[min., max.]	[-1.92, 9.32]	[-4.65, 5.77]	[-4.12, 6.35]
	red.		92.95%	89.56%
	M.S.E.	17.28	2.12	2.19
	cov.	32.75%	96.25%	95.25%
$n = 1000$	bias	3.84	0.30	0.38
	[min., max.]	[0.45, 8.10]	[-2.38, 3.36]	[-2.40, 3.63]
	red.		92.19%	90.10%
	M.S.E.	15.91	1.01	1.08
	cov.	6.25%	95.75%	95.75%

Anm.: ACE $= 0$, $X_1 \sim \mathcal{N}(0, 1^2)$, $\beta_1 = 10$, $\hat{e}(\vec{x}_i)_{\log}$: non-parametrisch geschätzter propensity score, $e(\vec{x}_i)$: tatsächlicher propensity score, red.: prozentuale bias-Reduktion, cov.: prozentuale coverage, [min., max.]: Minimum und Maximum der Punktschätzungen

Festzuhalten für eine Adjustierung mit $\hat{e}(\vec{x}_i)_{\log}$ ist eine mit dem festgelegten 95%−Niveau vereinbare coverage auch in großen Stichproben; trotz bestehender, geringfügiger Verzerrung und abnehmenden Schwankungen. Dementsprechend lässt sich für $n = 1000$ ein prozentuale Anteil von 95.75% an Stichproben-Konfidenzintervallen beobachten, die nach Adjustierung mit dem non-parametrisch geschätzten propensity score den ACE überdecken.

Tabelle 7 weist eindeutig auf, dass mit zunehmenden Stichprobenumfang die absoluten Abweichungen zwischen dem tatsächlichen $e(\vec{x}_i)$ sowie dem non-parametrisch geschätzten propensity score, $\hat{e}(\vec{x}_i)_{\log}$ abnehmen. Demnach näheren sich mit zunehmenden Stichprobenumfang beide Wahrscheinlichkeiten an.

4.2.3 Multivariate Normalverteilung

Tabelle 7: Vergleich beider geschätzten propensity scores

n	$\hat{E}(\mid e(\vec{x}_i) - \hat{e}(\vec{x}_i)_{\log} \mid)$
250	0.044
500	0.034
1000	0.026

In Anbetracht der resultierenden Verzerrungen bei der Schätzung des ACE nach Adjustierung mit den non-parametrisch geschätzten propensity score lässt sich bereits in kleinen Stichproben eine bias-Reduktion von fast 83% festhalten (vgl. Tabelle 8). Erstaunlicherweise nimmt der Wert zunächst mit zunehmenden Stichprobenumfang zu (ca. $85\%, n = 500$), während er mit $n = 1000$ abnimmt, so dass keine fortlaufende Zunahme der bias-Reduktion zum n erkennbar ist. Des weiteren ergeben sich bei der Adjustierung mit dem non-parametrisch geschätzten propensity score für alle Stichprobenumfänge bias-Reduktionen, die stets kleiner sind als die bias-Reduktionen, die einer Adjustierung mit dem parametrisch geschätzten propensity score folgen.

Während die Verzerrungen im Falle einer Adjustierung mit $e(\vec{x}_i)$ geringfügige Abweichungen vom ACE aufweisen, resultieren im Falle einer Adjustierung mit $\hat{e}(\vec{x}_i)_{\log}$ Verzerrungen, die stets im negativen Bereich liegen und auffällig sind. Damit weist diese Simulation gegenteilige Ergebnisse im Verhältnis zur ersten Simulation auf, wobei hervorgehoben werden muss, dass es sich zum Einen bei dem log-konkaven Dichteschätzer um ein non-parametrisches Schätzverfahren handelt, deren Konsistenzverhalten im Allgemeinen erst in großen Stichproben bemerkbar werden, zum Anderen, dass das treatment-assignment durch eine Interaktion zweier Kovariablen modelliert wurde und entsprechend eine Dichteschätzung erschwert wurde.

Es lässt sich festhalten, dass die mit dem non-parametrisch geschätzten propensity score adjustierten Punktschätzungen stets in einem größeren range liegen als die Punktschätzungen, die mit dem parametrisch geschätzten propensity score adjustiert werden. Während bei Adjustierung mit $e(\vec{x}_i)$ eine Abnahme des ranges bei zunehmenden Stichprobenumfang zu erkennen ist, lässt sich dies für die adjustierten Schätzungen mit $\hat{e}(\vec{x}_i)_{\log}$ nicht festhalten; stattdessen resultiert mit $n = 1000$ ein Intervall an Punktschätzungen, das größer ist als bei $n = 100$. Den bestehenden Verzerrungen und zunehmenden Intervallen an Punktschätzern entsprechend nehmen die M.S.E.-Schätzungen für die Adjustierung mit dem non-parametrisch geschätzten propensity score mit größer werdenden Stichprobenumfang zu, während diese bei Adjustierung

mit dem parametrisch-geschätzten propensity score abnehmen.

Tabelle 8: Ergebnisse der Simulation 2

		unadjustiert	$\hat{e}(\vec{x}_i)_{\log}$	$e(\vec{x}_i)$
$n = 250$	bias	7.90	-1.38	0.72
	[min., max.]	[2.91, 12.46]	[-8.52, 4.93]	[-5.39, 5.46]
	red.		82.53%	90.89%
	M.S.E.	65.40	6.75	4.22
	cov.	0.75%	99.5%	98.5%
$n = 500$	bias	7.96	-1.19	0.85
	[min., max.]	[3.23, 12.63]	[-6.92, 5.78]	[-5.08, 5.98]
	red.		85.05%	89.32%
	M.S.E.	65.82	5.56	3.75
	cov.	0%	98.5%	96.75%
$n = 1000$	bias	7.89	-1.59	0.83
	[min., max.]	[4.47, 11.19]	[-8.96, 6.73]	[-3.79, 4.44]
	red.		79.85%	89.48%
	M.S.E.	63.36	8.23	2.15
	cov.	0%	93%	96%

ACE $= 0$, $\vec{X} \sim \mathcal{N}(\vec{0}, \mathbf{\Sigma})$, $\beta_1 = 10$, $\hat{e}(\vec{x}_i)_{\log}$: non-parametrisch geschätzter propensity score, $e(\vec{x}_i)$: tatsächlicher propensity score, red.: prozentuale bias-Reduktion, cov.: prozentuale coverage, [min., max.]: Minimum und Maximum der Punktschätzungen

Die resultierenden coverages zeigen, dass der Anteil der Stichproben-Konfidenzintervallen, die nach Adjustierung mit dem non-parametrisch geschätzten propensity score den ACE überdecken, nicht mit festgelegten 95%−Niveau vereinbar ist. Während in kleinen Stichproben die coverages weit über dem 95%−Niveau liegen, liegt der prozentuale Anteil an geschätzten Intervallen, die den ACE beinhalten, in großen Stichproben unterhalb der 95%−Grenze. Für $n = 250$ sind die resultierenden Konfidenzintervalle in Abb.12 abgetragen. Es ist ersichtlich,

dass besonders breite Intervalle geschätzt werden, die auf entsprechend große geschätzte Varianzen des ACE zurückgeführt werden können. Tabelle 9 weist eindeutig einen auffälligen durchschnittlichen Abstand zwischen beiden geschätzten propensity scores auf, der nur geringfügig mit zunehmenden Stichprobenumfang abnimmt. Damit weisen die Ergebnisse dieser Simula-

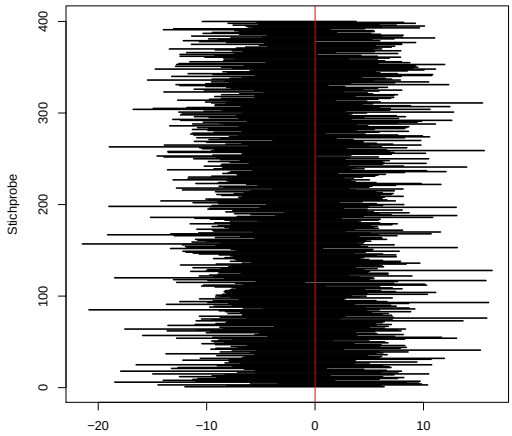

Abbildung 12: Konfidenzintervalle für den geschätzten ACE nach Adjustierung mit $\hat{e}(\vec{x}_i)_{\log}$

tionsstudie insgesamt Probleme bei der Adjustierung mit dem non-parametrisch geschätzten propensity score auf, die an späterer Stelle diskutiert werden.

Tabelle 9: Vergleich beider geschätzten propensity scores

n	$\hat{E}(\mid e(\vec{x}_i) - \hat{e}(\vec{x}_i)_{\log} \mid)$
250	0.185
500	0.176
1000	0.152

4.2.4 Student's $t-$Verteilung

Wie der Tabelle 10 zu entnehmen ist, weisen sowohl die mit dem non-parametrisch als auch mit dem parametrisch geschätzten propensity score adjustierten Punktschätzungen auffällige Verzerrungen über die Simulation hinweg auf.

Tabelle 10: Ergebnisse der Simulation 3

		unadjustiert	$\hat{e}(\vec{x}_i)_{\log}$	$e(\vec{x}_i)$
$n = 250$	bias	8.69	1.98	2.13
	[min., max.]	[-0.21, 17.33]	[-5.06, 9.36]	[-5.41, 9.79]
	red.		77.21%	75.48%
	M.S.E.	84.69	9.69	10.15
	cov.	13.5%	91%	94%
$n = 500$	bias	8.80	2.22	2.29
	[min., max.]	[3.42, 15.53]	[-2.11, 9.10]	[-2.12, 7.96]
	red.		74.77%	73.98%
	M.S.E.	81.12	7.53	7.75
	cov.	0.75%	79.5%	88.25%
$n = 1000$	bias	8.78	2.21	2.27
	[min., max.]	[4.38, 12.65]	[-0.93, 5.89]	[-0.93, 5.94]
	red.		74.83%	74.15%
	M.S.E.	79.17	6.28	6.46
	cov.	0%	60%	72.5%

ACE $= 0$, $X_1 \sim T(3)$, $\beta_1 = 10$, $\hat{e}(\vec{x}_i)_{\log}$: non-parametrisch geschätzter propensity score, $e(\vec{x}_i)$: tatsächlicher propensity score, red.: prozentuale bias-Reduktion, cov.: prozentuale coverage, [min., max.]: Minimum und Maximum der Punktschätzungen

Damit scheint die Wahl einer heavy-tailed verteilten Störvariablen, die im Verhältnis zu einer normalverteilten Zufallsvariablen extremere Realisierungen x_i annehmen kann, zu Problemen bei der Adjustierung mit dem geschätzten propensity score zu führen, die nicht direkt auf

eine der Schätzmethoden für den propensity score zurückgeführt werden kann, da die resultierenden Verzerrungen beider Schätzmethoden für alle n direkt vergleichbar sind. Trotz der auffälligen Verzerrungen weisen beide propensity score-Adjustierungen bei Schätzung des ACE eine Reduktion von mindestens 74% desjenigen bias auf, der auf die konfundierende Störvariable zurückgeführt werden kann. Nach Adjustierung mit dem non-parametrisch geschätzten propensity score resultieren stets größere bias-Reduktionen als mit dem parametrisch geschätzten.

In Anbetracht bestehender Verzerrungen und mit zunehmenden Stichprobenumfang abnehmenden Intervallen an Punktschätzungen nehmen die coverages nach Adjustierung mit beiden propensity scores mit zunehmender Stichprobengröße ab. Für $n = 250$ werden die Konfidenzintervalle nach Adjustierung mit $\hat{e}(\vec{x}_i)_{\log}$ nicht mehr so geschätzt, dass die resultierende coverage vereinbar mit einem Niveau von 95% ist, mit $n = 500$ gilt dies auch für eine Adjustierung mit dem tatsächlichen propensity score, $e(\vec{x}_i)$. Zurückzuführen ist dieses Ergebnis eindeutig auf die Tendenz, dass mit einer der beiden Adjustierungen stets eine Verzerrung einhergeht, während gleichzeitig mit zunehmenden n die Intervalle an möglichen Punktschätzungen abnehmen. Dem entsprechend lässt sich eine zum zunehmenden n einhergehende Abnahme der M.S.E.-Schätzungen festhalten. Tabelle 11 weist auf, dass bereits mit $n = 250$ geringe Differenzen zwischen den beiden geschätzten propensity scores vorliegen, die mit zunehmenden Stichprobenumfang abnehmen. Mit $n = 500$ sind die Differenzen zwischen beiden propensity score-Schätzungen geringer als bei Vorliegen einer univariat-normalverteilten Kovariablen (vgl. Tabelle 7).

Tabelle 11: Vergleich beider geschätzten propensity scores

n	$\hat{E}(\mid e(\vec{x}_i) - \hat{e}(\vec{x}_i)_{\log} \mid)$
250	0.042
500	0.032
1000	0.022

4.2.5 Pareto-Verteilung

Die in Tabelle 12 dargelegten Ergebnisse weisen auf, dass trotz einer Verletzung der log-Konkavität bei Vorliegen einer Pareto-verteilten Störvariablen eine Adjustierung mit dem non-parametrisch geschätzten propensity score zu einer erheblichen Reduktion des bias führt, der ohne Adjustierung auf die konfundierende Störvariable zurückgeführt werden kann. Mit $n = 250$

resultiert eine bias-Reduktion von ca. 80%, die jedoch mit zunehmenden n abnimmt.

Tabelle 12: Ergebnisse der Simulation 4

		unadjustiert	$\hat{e}(\vec{x}_i)_{\log}$	$e(\vec{x}_i)$
$n = 250$	bias	5.22	1.06	1.15
	[min., max.]	[-13.89, 17.08]	[-10.09, 11.07]	[-10.51, 12.88]
	red.		79.69%	77.97%
	M.S.E.	45.43	16.90	18.29
	cov.	87.5%	98.25%	97.25%
$n = 500$	bias	5.06	1.24	1.62
	[min., max.]	[-2.09, 13.02]	[-7.39, 9.94]	[-5.77, 10.32]
	red.		75.49%	67.98%
	M.S.E.	34.04	11.02	12.07
	cov.	73.75%	97.5%	99.75%
$n = 1000$	bias	5.22	1.49	1.76
	[min., max.]	[-0.06, 11.19]	[-5.60, 8.99]	[-5.17, 8.99]
	red.		71.46%	66.28%
	M.S.E.	31.23	6.70	7.23
	cov.	43.25%	95.75%	100%

ACE $= 0$, $X_1 \sim Par(5,5)$, $\beta_1 = 10$, $\hat{e}(\vec{x}_i)_{\log}$: non-parametrisch geschätzter propensity score, $e(\vec{x}_i)$: tatsächlicher propensity score, red.: prozentuale bias-Reduktion, cov.: prozentuale coverage, [min., max.]: Minimum und Maximum der Punktschätzungen

Für die, sowohl mit dem non-parametrisch als auch mit dem parametrisch geschätzten propensity score adjustierten Punktschätzungen lassen sich für alle n Verzerrungen festhalten, die mit zunehmenden Stichprobenumfang zunehmen; die mit dem theoretischen propensity score, $e(\vec{x}_i)$, adjustierten Punktschätzungen weisen für alle n stets eine größere Verzerrung auf als nach Adjustierung mit dem non-parametrisch geschätzten propensity score.

In Anbetracht der resultierenden Überdeckungsraten führt die Adjustierung einer Pareto-verteilten Störvariablen zu Problemen bei der Varianzschäzung, irrelevant, ob die Schätzung

des ACE mit dem parametrisch oder non-parametrisch geschätzten propensity score adjustiert wird: Für $n = 250$ resultieren in beiden Fällen coverages, die weit über dem festgelegten 95% liegen. Während mit zunehmenden Stichprobenumfang die Überdeckungsraten gegeben einer Adjustierung mit dem non-parametrisch geschätzten propensity score abnehmen und mit $n = 1000$ approximativ dem Niveau von 95% entsprechen, nehmen die Überdeckungsraten gegeben einer Adjustierung mit dem parametrisch geschätzten propensity score mit zunehmenden Stichprobenumfang zu, so dass mit $n = 1000$ eine coverage von 100% resultiert.

In Abbildung 13 sind die resultierenden Stichproben-Konfidenzintervalle für $n = 250$ sowohl nach Adjustierung mit dem non-parametrisch wie auch mit dem parametrisch geschätzten propensity score abgetragen. Hier ist die Tendenz verdeutlicht, dass gegeben einer Adjustierung mit dem parametrisch geschätzten propensity score breitere Intervalle geschätzt werden als mit dem non-parametrisch geschätzten propensity score.

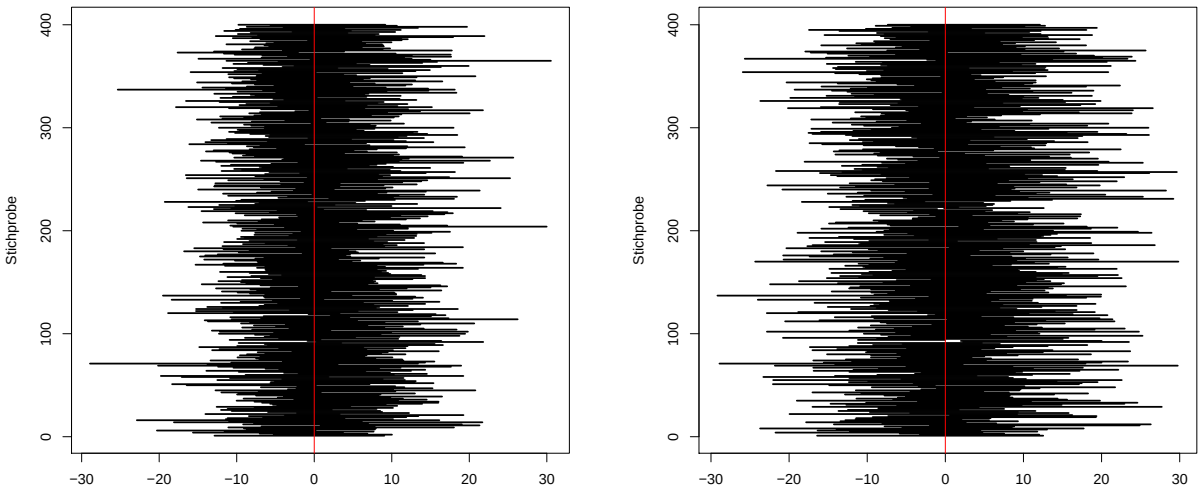

Abbildung 13: Konfidenzintervalle bei $n = 250$ für den geschätzten ACE nach Adjustierung mit beiden propensity scores, links: non-parametrisch geschätzter propensity score

Beide propensity score-Adjustierungen führen mit zunehmenden Stichprobenumfang zu einer Abnahme der M.S.E.-Schätzungen; mit stets größeren Verzerrungen bei Adjustierung mit dem parametrisch geschätzten propensity score sind die resultierenden M.S.E.-Werte für jedes n entsprechend größer als mit dem non-parametrisch geschätzten propensity score.

Tabelle 13: Vergleich beider geschätzten propensity scores

n	$\hat{E}(\mid e(\vec{x}_i) - \hat{e}(\vec{x}_i)_{\log} \mid)$
250	0.017
500	0.011
1000	0.008

Tabelle 13 weist auf, dass bereits in kleinen Stichproben nur geringe Abweichungen zwischen den beiden geschätzten propensity scores vorliegen, die mit größer werdenden Stichprobenumfang abnehmen.

4.3 Diskussion

Grundlegend zeigen die Simulationen auf, dass die Verwendung des vorgeschlagenen, non-parametrisch geschätzten propensity scores in Konsequenz zu einer Reduktion des bias führt, der einer unadjustierten Schätzung des ACE folgt. Somit kann basal das Ziel, das bei Verwendung der propensity score-Methoden verfolgt wird, sichergestellt werden: bedingt an $e(\vec{x}_i)$ soll die Konfundierung der in $\vec{x}_i$ erfassten Kovariablen behoben und eine konsistente Schätzung des ACE ermöglicht werden. Dementsprechend stellen die resultierenden bias-Reduktionen empirische Indizien dafür dar, dass ein Großteil der Konfundierung durch die Störvariablen behoben werden kann.

An dieser Stelle muss darauf verwiesen werden, dass ein Anteil der resultierenden Verzerrungen nach entsprechender Adjustierung stets auf die Methodik der Quintil-Stratifikation zurückgeführt werden kann: Es zeigt sich, mit Ausnahme der zweiten Simulationsstudie, dass die Verzerrungen, die einer Adjustierung mit dem non-parametrisch geschätzten propensity score folgen, direkt vergleichbar mit denen der Adjustierung mit dem parametrisch geschätzten propensity score sind und damit eindeutig in einen direkten Zusammenhang mit der Adjustierungmethode gebracht werden können. Wie bereits theoretisch begründet, handelt es sich bei der verwendeten propensity score-Stratifikation um ein Adjustierungsverfahren, das zu inkonsistenten Ergebnissen bei der Schätzung des ACE führt, da die resultierenden Varianzen des propensity scores innerhalb der strata mit bestehenden Konfundierungen durch die Störvariablen einhergehen. Dementsprechend führt eine Schätzung des ACE nach Quintil-Stratifikation grundlegend zu verzerrten Schätzergebnissen und die entsprechenden Verzerrungen nehmen nicht mit zunehmenden Stichprobenumfang ab. Damit scheint eine Begründung für die Wahl der Quinitl-Stratifikation als benutztes Adjustierungsverfahren notwendig: zum Einen lässt sich aufführen, dass bei Verwendung einer der weighting-Methoden im Zusammenhang mit der non-

parametrischer Schätzung des propensity scores Probleme bei der Varianzschätzung auftreten können, da propensity scores im Wertebereich $[0, 1]$ resultieren können. Zum Anderen handelt es sich hinsichtlich der langen Verarbeitungsdauer bei Verwendung des non-parametrischen Dichteschätzers bei der Quintil-Stratifikation um ein Adjustierungsverfahren, das sich entsprechend schnell und ohne großen Verarbeitungsaufwand umsetzen lässt.

Insbesondere die zweite Simulation, die eine multivariate Normalverteilung der fünf Störvariablen als Verteilungsmodell im datengenerierenden Prozess modelliert, weist Probleme bei der non-parametrischen Schätzung des propensity scores auf, die in einen Zusammenhang mit der *curse of dimensionality* gebracht werden können: während die Schätzungen des ACE nach Adjustierung mit dem parametrisch geschätzten propensity score geringfügige und positive Verzerrungen aufweisen, resultieren nach Adjustierung mit dem non-parametrisch geschätzten propensity score negative und betragsmäßig größere Verzerrungen. Für alle n lässt sich nach Adjustierung mit den non-parametrisch geschätzten propensity score stets ein breiteres Intervall an Punktschätzungen festhalten als nach Adjustierung mit dem parametrisch geschätzten propensity score. Zusätzlich resultiert für $n = 1000$ ein größerer range an Punktschätzungen als für $n = 250$, während nach parametrischer Schätzung des propensity scores die resultierenden Intervalle für den geschätzten ACE abnehmen. An dieser Stelle muss jedoch betont werden, dass die zu schätzende Selektionswahrscheinlichkeit abhängig von einer Interaktion zweier Kovariablen ist und die Schätzung der entsprechenden Dichte verkompliziert wird. Demnach müssten in diesem Fall eindeutig größere Stichproben als die verwendeten Stichprobengrößen gezogen werden, um prüfen zu können, ob entsprechende Verzerrungen und breite Intervalle an Punktschätzungen mit zunehmenden n abnehmen und entsprechende Probleme, die mit der Dimensionalität und Komplexität der Dichte in Beziehung stehen, behoben werden können.

Auch die Verwendung einer heavy tailed-verteilten Störvariablen in der dritten Simulation führt zu Problemen bei der Adjustierung, die anteilsmäßig auf die Methodik der Quintil-Stratifikation zurückgeführt werden kann. Für alle n sind die Verzerrungen nach Adjustierung mit beiden geschätzten propensity scores vergleichbar und sind zurückzuführen auf mögliche Extremwerte bei Verwendung der Student's $t-$verteilten Störvariable, die in Konsequenz zu kleinen bzw. großen geschätzten Selektionswahrscheinlichkeiten führen. Entsprechend nehmen die Varianzen der geschätzten propensity scores in den strata und damit die bestehenden Konfundierungen zu. Dieser Effekt schlägt sich in entsprechenden Intervallschätzungen nie-

der, wie die resultierenden Überdeckungsraten mit zunehmenden Stichprobenumfang aufweisen: Im Falle beider Adjustierungen resultieren Verzerrungen und abnehmende Intervalle an Punktschätzungen, so dass die geschätzte Konfidenzintervalle schmaler werden und den ACE anteilsmäßig seltener überlagern. An dieser Stelle wäre es notwendig, zu prüfen, ob dieser Effekt mit zunehmender Anzahl an Quantilen, die zur propensity score-Stratifikation genutzt werden, oder mit einem exakten matching in großen Stichproben entsprechend abnimmt.

Erstaunlicherweise zeigt die vierte Simulation, dass der log-konkave Dichteschätzer bei Verletzung der log-Konkavität robust reagiert: Gegeben einer Pareto-verteilten Kovariable, die als konfundierende Störvariable genutzt wurde und keine log-konkave Dichtefunktion besitzt, zeigt sich, dass eine Adjustierung mit dem propensity score, der die geschätzten, bedingten Dichten nutzt, zu vergleichbaren Ergebnissen führt wie nach Adjustierung mit dem parametrisch geschätzten propensity score. Für alle Stichprobenumfänge sind die resultierenden Verzerrungen bei der Schätzung des ACE nach entsprechender Adjustierung direkt vergleichbar und weisen die selbe Tendenz auf, dass mit zunehmenden n die resultierenden Verzerrungen geringfügig zunehmen. Insgesamt resultieren bei Adjustierung mit dem non-parametrisch geschätzten propensity score größere Intervalle an möglichen Punktschätzungen als nach Adjustierung mit dem parametrisch geschätzten propensity score, so dass eine non-parametrische Schätzung damit zu größeren Unsicherheiten bei Punktschätzung des ACE führt. Bei Vergleich beider Adjustierungen zeigt sich, dass insbesondere hinsichtlich der coverage gegenläufige Effekte festzustellen sind: Während einer Adjustierung mit dem parametrisch geschätzten propensity score bei zunehmenden Stichprobenumfang derartig breite Intervallschätzungen folgen, dass mit $n = 1000$ alle geschätzten Intervalle den ACE überlagern, nehmen die resultierenden Überdeckungsraten bei Adjustierung mit dem non-parametrisch geschätzten propensity score mit zunehmenden n ab und sind erst mit $n = 1000$ vereinbar mit dem $95\%-$Niveau.

Trotz der Ergebnisse der vierten Simulation sollte keine Verallgemeinerung auf weitere Dichten, die nicht log-konkav sind, impliziert werden: insbesondere in den Human- und Sozialwissenschaften werden zugrunde liegende empirische Phänomene teilweise mit Verteilungsannahmen modelliert, deren Dichten nicht log-konkav sind; Beispiele hierfür wurden in der Einleitung aufgeführt. Inwieweit sich die Ergebnisse der vierten Simulation, die eine Pareto-Verteilung zu Grunde legte, übertragen lassen auf mögliche χ^2- oder $F-$Verteilungen, kann an dieser Stelle nicht geklärt werden.

5 Adjustierte Schätzung des ACE bei latenten Störvariablen

5.1 Einleitung

Sämtliche der bisherig eingeführten Adjustierungsmethoden zur Schätzung des ACE basierten auf einer messfehlerfreien Erhebung der Werte $\vec{x}_i$, so dass *manifeste*, d.h. empirisch direkt zugängliche und fehlerfrei gemessene Störvariablen unterstellt worden sind, die sowohl die treatment-Selektion wie auch den response-Wert der statistischen Einheit i bedingen. Es lässt sich jedoch einwenden, dass insbesondere in den Human- und Sozialwissenschaften nicht-randomisierte treatment-assignments vorliegen, die auf *latenten Kovariablen*, d.h. auf Konstrukten, die nicht vollständig reliabel durch manifeste Skalen erfasst werden, beruhen. In Anlehnung an Rubin et al. (2004) lässt sich als Beispiel eine Selektion zu einem Matheförderkurs als treatment in Abhängigkeit von beobachteten Mathetestwerten, die damit Indikatoren für eine ggf. förderungsbedürftige *generelle Mathefähigkeit* darstellen, nennen.

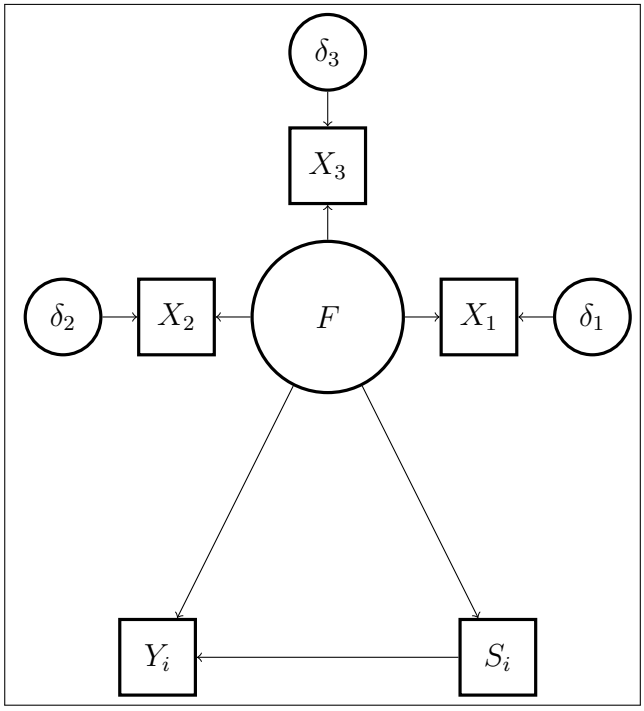

Abbildung 14: Annahme einer latenten Störvariablen

Dem Beispiel folgend lässt sich damit unterstellen, dass eine latente Variable, auch *Faktor* oder *Konstrukt* bezeichnet, vorliegt, die einerseits in ihrer möglichen Wirkung als Störvariable die Entscheidung, in welche Experimentalbedingung die Einheit i selegiert wird, sowie die potentiellen response-Werte beeinflusst und die andererseits als latentes Merkmal nicht vollständig reliabel durch die Variablen $X_1, \ldots, X_p$ erfasst wird. In messtheoretischer Hinsicht wird folgend damit ein Unterschied bezüglich der potentiellen Störvariablen als latentes und empirisch nicht direkt zugängliches Merkmal, und den manifest erhobenen Kovariablen $\vec{X}^t = (X_1, \ldots, X_p)$ getroffen, die als Indikatoren den Faktor messfehlerbelastet abbilden. Abbildung 14 verdeutlicht die Idee der latenten Störvariablen samt manifesten Indikatorva-

riablen, die den Faktor mit einem entsprechenden Fehler, δ_j, $j = 1, \ldots, p$, messen.

Insbesondere die Psychologie als empirische Wissenschaft ist konfrontiert mit dem Problem der latenten Variablen, so dass es naheliegend erscheint, nach Adjustierungsmöglichkeiten zur Schätzung des ACE bei einer treatment-Selektion in Abhängigkeit von einem latenten Konstrukt, welches durch entsprechende Indikatorvariablen messfehlerbehaftet erfasst wird, zu suchen.

Die Notwendigkeit, den auftretenden Messfehler bei der Schätzung eines Regressionskoeffizientens β, beispielsweise im linearen Regressionsmodell bei der Schätzung des ACE, konstant zu halten, lässt sich leicht durch das *errors-in-variables*-Modell zeigen (z.B., Chesher, 1991; Yuan & Bentler, 2007).

5.1.1 Exkurs: Errors-In-Variables-Problem

Angenommen, es interessiert das lineare Regressionsmodell

$$y_i = \alpha + \beta \cdot x_i^* + \epsilon_i, \quad i = 1, \ldots, n,$$

wobei x_i^* ein latenter Wert ist und durch

$$x_i = x_i^* + \eta_i,$$

mit Gültigkeit der Annahmen $Cov(\eta_i, x_i^*) = 0, E(\eta_i) = 0, Var(\eta_i) = \sigma_\eta^2$ erfassbar ist.

Ein Schätzmodell, das die y_i auf die entsprechend beobachteten x_i regressiert, d.h.

$$y_i = \hat{\alpha} + \hat{\beta} \cdot x_i + \hat{\epsilon}_i$$

hätte eine inkonsistente Schätzung des interessierenden Parameters β zur Folge, wie sich durch Einsetzen und Umformen leicht zeigen lässt:

$$\hat{\beta} \xrightarrow{f.s.} \frac{Cov(X,Y)}{Var(X)} = \frac{\beta \cdot \sigma_{X^*}^2}{\sigma_{X^*}^2 + \sigma_\eta^2} = \frac{\beta}{1 + \frac{\sigma_\eta^2}{\sigma_{X^*}^2}}.$$

Damit folgt im Grenzwert mit $\sigma_\eta^2 > 0$ und $\sigma_{X^*}^2 > 0$ eine Unterschätzung des Effektes β, wenn dieser ausgehend von den beobachteten Werten x_i geschätzt wird (*attenuation bias*).

5.1.2 Exkurs: Lineare Strukturgleichungsmodelle

Eine insbesondere in den Sozialwissenschaften populäre Möglichkeit, Regressionskoeffizienten unter Konstanthaltung möglicher Messfehler, die sich bei Messung eines latenten Merkmals

ergeben, zu schätzen, stellt die Spezifikation von *linearen Strukturgleichungsmodellen (SEM oder LISREL)* dar, die in einer Vielzahl von Veröffentlichungen Anwendung finden (z.B., Anderson & Gerbing, 1988; Bentler, 1980; Burns & Nolen-Hoeksema, 1992; Phillip, Gus, Rodney & John, 2003; Sawyer, 1992; Semmer, Tschan, Meier, Facchin & Jacobshagen, 2010).

Die Popularität von Strukturgleichungsmodellen lässt sich zurückführen auf eine mögliche Spezifikation von *Strukturmodellen*, innerhalb derer die in das Modell aufgenommenen Variablen als exogen und endogen klassifiziert werden und sich entsprechende Effekte, die die exogenen Variablen auf die endogenen Variablen ausüben, schätzen lassen. Die Besonderheit der Strukturgleichungsmodellierung ergibt sich durch eine mögliche Erweiterung der Strukturmodelle mit einer zusätzlichen Spezifikation von *Messmodellen*, innerhalb derer latente Variablen durch manifest erhobene Variablen samt Messfehler operationalisiert und dem folgend in das Strukturmodell aufgenommen werden können. Damit ermöglicht die Formulierung eines linearen Strukturgleichungsmodells die Schätzung der Effekte exogener latenter Variablen auf endogene latente Variablen unter Konstanthaltung der jeweiligen Messfehler, die sich bei Modellierung der latenten Variablen einstellen.

Somit ist es in einem Strukturgleichungsmodell möglich, die Effekte exogener auf endogene Variablen zu schätzen, wobei die spezifizierten Modelle vollständig auf manifesten Variablen beruhen können (Ansatz der *Pfadanalyse* als Verallgemeinerung der *multiplen Regressionsanalyse* mit mehreren abhängigen Variablen) oder auch latente Variablen beinhalten können, die durch entsprechende Indikatorvariablen operationalisiert werden (Ansatz der *konfirmatorischen Faktorenanalyse*) und somit eine Schätzung der interessierenden Regressionskoeffizienten des Strukturmodells unter Konstanthaltung der Messfehler, die bei Operationalisierung der latenten Variablen auftreten, ermöglichen.

Abbildung 15 gibt ein Beispiel für ein SEM, dargestellt in Form eines *nicht-rekursiven Pfaddiagramms*, innerhalb dessen die latente Variable ξ_1, operationalisiert durch die manifesten Variablen X_1 und X_2, im Strukturmodell die exogene Variable hinsichtlich der latenten Variablen η_1 und η_2, operationalisiert durch die manifesten Variablen $Y_1, \ldots, Y_4$, darstellt. In diesem Strukturmodell lässt sich die endogene latente Variable η_1 als eine weitere exogene Variable mit einer Wirkung auf die latente Variable η_2 klassifizieren; der gerichtete Effekt, den η_1 auf η_2 ausübt, wird durch β_1 indiziert. Die Fehlerterme $\vec{\delta}$ sowie $\vec{\epsilon}$ entsprechen in diesem Modell den Messfehlern, die sich einstellen, wenn eine manifeste Variable als Indikator für eine

latente Variable benutzt wird, die Fehlerterme $\vec{\zeta}$ entsprechen den Regressionsfehlern, die zur Modellierung der endogenen Variablen durch die exogenen Variablen benötigt werden.

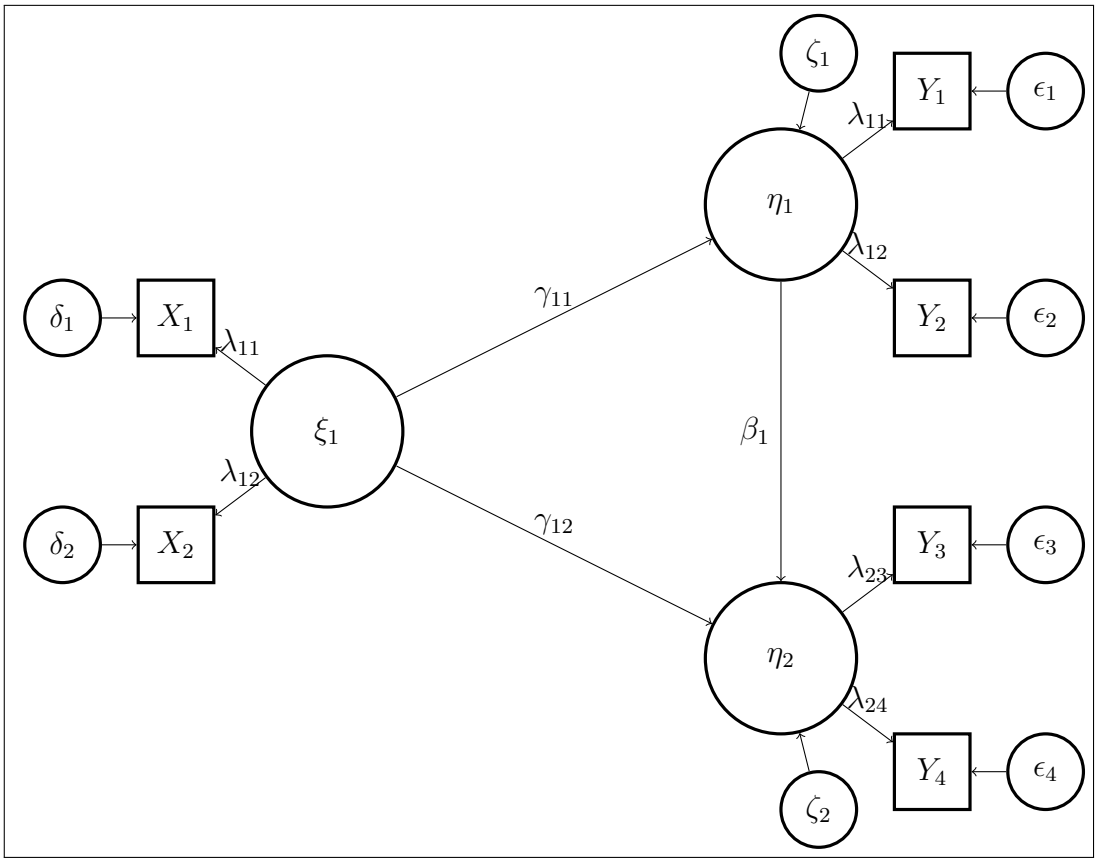

Abbildung 15: Beispiel für ein Strukturgleichungsmodell

Die dem Diagramm resultierenden Pfade lassen sich in lineare Gleichungssysteme überführen, deren unbekannten Parameter anhand der Daten geschätzt werden sollen (Mulaik, 2009). Das obige Strukturmodell lässt sich in die folgenden Gleichungssysteme überführen:

$$\begin{array}{ccccccc}
\eta_1 & = & & & \gamma_{11} \cdot \xi_1 & + & \zeta_1 \\
\eta_2 & = & \beta_1 \cdot \eta_1 & + & \gamma_{12} \cdot \xi_1 & + & \zeta_2
\end{array}$$

<u>in allgemeiner Darstellung:</u>

$$\vec{\eta} = \mathbf{B}\vec{\eta} + \mathbf{\Gamma}\vec{\xi} + \vec{\zeta}$$

Das Messmodell der exogenen Variablen lässt sich analog durch folgende Gleichungssysteme angeben:

$$
\begin{array}{ccccccc}
X_1 & = & \mu_{X_1} & + & \lambda_{11} \cdot \xi_1 & + & \delta_1 \\
X_2 & = & \mu_{X_2} & + & \lambda_{12} \cdot \xi_1 & + & \delta_2
\end{array}
$$

<u>in allgemeiner Darstellung:</u>

$$
\vec{X} \quad = \quad \vec{\mu}_{\vec{X}} \quad + \quad \mathbf{\Lambda}_{\vec{X}} \vec{\xi} \quad + \quad \vec{\delta},
$$

das Messmodell der endogenen Variablen (in analoger und allgemeiner Darstellung) durch:

$$
\vec{Y} \quad = \quad \vec{\mu}_{\vec{Y}} \quad + \quad \mathbf{\Lambda}_{\vec{Y}} \vec{\eta} \quad + \quad \vec{\epsilon}.
$$

Die Ermittlung eindeutiger Schätzwerte für den Parametervektor $\vec{\theta}$, der die $q-$unbekannten Parameter des Modells beinhaltet, sowie eine Testung des Schätzmodells hinsichtlich der Plausibilität des spezifizierten Modells an die beobachteten Daten (sog. *goodness-of-fit*-Test) ist gebunden an Restriktionen, die zu der Annahme eines *identifizierten Modells* führen (z.B., Anderson & Gerbing, 1988; Fox, 2006; Mulaik, 2009; Yuan & Bentler, 2007):

1. Das spezifizierte Modell gilt erst als identifiziert, wenn zusätzlich zu den interessierenden Regressionskoeffizienten, zusammengefasst in den Matrizen $\mathbf{\Gamma}$ und $\mathbf{B}$, sowie den Faktorladungen, zusammengefasst zur Matrix $\mathbf{\Lambda}$, sämtliche unbekannte (Ko-)Varianzen in das zu schätzende Modell aufgenommen und ggf. fixiert werden. Hierzu zählt die Spezifikation der Diagonalmatrix $\mathbf{\Theta}^2_\epsilon$, die die Varianzen der Messfehler von $\vec{X}$, $Var(\vec{\epsilon})$, zusammenfasst, die Spezifikation der Diagonalmatrix $\mathbf{\Theta}^2_\delta$, die die Varianzen der Messfehler von $\vec{Y}$, $Var(\vec{\delta})$, zusammenfasst, sowie die Spezifikation der Varianz-Kovarianzmatrix $\mathbf{\Psi}$ mit den (Ko-)Varianzen der Regressionsfehler $\vec{\zeta}$ sowie der Varianz-Kovarianzmatrix der latenten exogenen Variablen $\vec{\xi}$, $\mathbf{\Phi}$.

   Insbesondere eine Fixierung der Varianz der latenten exogenen Variablen, $Var(\xi)$, erscheint hinsichtlich der Operationalisierung durch die Indikatorvariablen notwendig, damit eine entsprechende *Skalierung* bzw. *Maßeinheit* vorliegt, mit der das Konstrukt gemessen wird. Zumeist wird die Varianz der latenten exogenen Variablen mit $Var(\xi) = 1$ fixiert (die Diagonale der Matrix $\mathbf{\Phi}$ besteht entsprechend vollständig aus 1.0-Einträgen)

oder sie lässt sich schätzen in Folge einer Fixierung einer der Faktorladungen zu $\lambda = 1$, so dass die Varianz der latenten Variablen festgelegt wird durch die Varianz der entsprechenden Indikatorvariablen (Yuan & Bentler, 2007). Zusätzlich gilt als Voraussetzung zur Identifikation des Modells, dass mögliche Kovarianzen zwischen den Messfehlern formuliert werden, wenn diese im Modell entsprechend berücksichtigt werden sollen.

2. Wie folgend zu zeigen ist, stehen die unbekannten Parameter des linearen Strukturgleichungsmodells in einer eindeutigen Beziehung mit der Varianz-Kovarianzmatrix $\mathbf{\Sigma}$. Das Identifikationsproblem formuliert die Restriktion, dass die Parameter des Modells anhand beobachteter Daten nur schätzbar sind, wenn das Modell nicht *unteridentifiziert* ist. Demnach darf das Schätzmodell nicht mehr unbekannte Parameter beinhalten als Möglichkeiten zur Schätzung dieser bestehen. Präziser lässt sich festhalten: Mit p manifesten Variablen stehen insgesamt $p-$Mittelwerte sowie $\frac{p(p+1)}{2}-$Einträge in der Varianz-Kovarianzmatrix fest; die Schätzung von q unbekannten Parametern kann nur erfolgen, wenn $q \leq \frac{p(p+3)}{2}$ (vgl. Yuan & Bentler, 2007). Im Falle $q = \frac{p(p+3)}{2}$ reproduziert das Modell die Varianz-Kovarianzmatrix vollständig und es kann kein Modelltest, der die Anpassungsgüte des Modells an die Daten prüft, stattfinden. Die Parameter eines solchen Modells sind damit eindeutig identifiziert; ein solches Modell entspricht dem *gesättigten Modell*.

3. Yuan und Bentler (2007) zeigen, dass, von wenigen stark restriktiven Ausnahmen abgesehen, einer latenten Variable im Messmodell mindestens drei Indikatorvariablen zugeordnet sein müssen, damit das spezifizierte Messmodell identifiziert ist.

Die Schätzung der unbekannten Modellparameter $\vec{\theta}$ kann durch verschiedene Schätzmethoden erfolgen, am gängigsten ist die Maximum-Likelihood-Schätzung, die von Jöreskog (1970) vorgeschlagen wurde und eine multivariate Normalverteilung der manifesten Variablen $\vec{X}^t = (\vec{X}, \vec{Y})$, d.h. $\vec{X} \sim \mathcal{N}(\vec{\mu}, \mathbf{\Sigma})$, voraussetzt: Mit Gültigkeit der Annahmen $E(\vec{\eta}) = E(\vec{\xi}) = E(\vec{\zeta}) = E(\vec{\epsilon}) = 0$ sowie der linearen Unabhängigkeit der Fehlerterme von den Variablen des Mess- und Strukturmodells lässt sich zeigen, dass die Varianz-Kovarianzmatrix $\mathbf{\Sigma}$ der manifesten Variablen durch die unbekannten Parameter des Modells, $\vec{\theta}$, dargestellt werden kann (sog. *model implied* Varianz-

Kovarianzmatrix, $\mathbf{\Sigma}(\vec{\theta})$); es gilt:

$$\begin{aligned}\mathbf{\Sigma} &= \begin{pmatrix} \mathbf{\Sigma}_{YY} & \mathbf{\Sigma}_{XY} \\ \mathbf{\Sigma}_{YX} & \mathbf{\Sigma}_{XX} \end{pmatrix} \\ &= \begin{pmatrix} \mathbf{B}^{-1}\mathbf{\Gamma}\mathbf{\Phi}\mathbf{\Gamma}^t\mathbf{B}^{t^{-1}} + \mathbf{B}^{-1}\mathbf{\Psi}\mathbf{B}^{t^{-1}} + \mathbf{\Theta}_\epsilon^2 & \mathbf{B}^{-1}\mathbf{\Gamma}\mathbf{\Phi} \\ \mathbf{\Phi}\mathbf{\Gamma}^t\mathbf{B}^{t^{-1}} & \mathbf{\Phi} + \mathbf{\Theta}_\delta^2. \end{pmatrix}.\end{aligned}$$

Durch Anwendung des *Fletcher-Powell*-Algorithmus lassen sich iterativ Punktschätzungen für $\vec{\theta}$ bestimmen (Maximum-Likelihood-Schätzungen), die die log-Likelihood-Funktion

$$\log(\mathcal{L}) = -\frac{n}{2} \cdot \left[\log \mid \mathbf{\Sigma}(\vec{\theta}) \mid + \mathrm{tr}\left(\mathbf{\Sigma}(\vec{\theta})^{-1}\mathbf{S}\right)\right]$$

maximieren und entsprechend die ML-Diskrepanzfunktion,

$$D_{ML} = \log \mid \mathbf{\Sigma}(\vec{\theta}) \mid - \log \mid \mathbf{S} \mid + \mathrm{tr}\left[\mathbf{S}\mathbf{\Sigma}(\vec{\theta})^{-1}\right] - p,$$

minimieren, die in einem unmittelbaren Zusammenhang mit der Likelihood-Quotienten-Statistik steht (Yuan & Bentler, 2007); es gilt:

$$T_{ML} = n \cdot D_{ML}.$$

Unter der Annahme multivariat-normalverteilter manifester Variablen gilt für diese Statistik asymptotisch $T_{ML} \overset{a}{\sim} \chi^2_{df}$, wobei $df = \frac{p(p+3)}{2} - q$. Somit lässt sich das spezifizierte Modell mit den resultierenden Schätzungen durch einen *goodness-of-fit*-Test prüfen, der mit hohen χ^2-Werten einhergeht, wenn die Modellstruktur nicht mit der beobachteten Kovarianzmatrix übereinstimmt. Dem entsprechend lässt sich mit diesem χ^2-Test die Nullhypothese, $H_0 : \mathbf{\Sigma} = \mathbf{\Sigma}(\vec{\theta})$, testen.

Der aufgeführten Maximum-Likehood-Schätzung stehen weitere Verfahren der Parameterschätzung gegenüber, beispielsweise Bayes-Methoden (z.B., Muthén & Asparouhov, 2011; Rabe-Hesketh, Skrondal & Zheng, 2007; Scheines, Hoijtink & Boomsma, 1999), Kleinste-Quadrate-Methoden (Olsson, Foss, Troye & Howell, 2000) oder asymptotisch verteilungsfreie Methoden (*ADF-Methoden*, s. z.B. Browne, 1984), die von Normalverteilungsannahmen absehen und entsprechend abweichende Diskrepanzfunktionen definieren. Für einen Überblick über diese Schätzmethoden bieten sich die entsprechenden Kapitel von Mulaik (2009) sowie Bollen (1989) an.

5.1.3 Observational Studies bei latenten Variablen

Der Zusammenhang zur unkonfundierten Schätzung des ACE mit dem Problem der latenten Variablen lässt sich unmittelbar herstellen: Unter der Annahme, dass die latente Variable ξ eine Störvariable darstellt, lässt sich das logistische Modell, das der treatment-Selektion zumeist als Modell zu Grunde gelegt wird, durch

$$\Pr(S_i = 1 \mid \xi_i) = \frac{1}{1 + \exp(-(\gamma_0 + \gamma_1 \cdot \xi_i))} \tag{48}$$

darstellen, das zu Grunde liegende response-Modell durch das lineare Modell

$$Y_i = \alpha + \tau \cdot S_i + \gamma_2 \cdot \xi_i + \zeta_i, \qquad \alpha = E(Y_i \mid S_i = 0) = E(Y_0). \tag{49}$$

Eine konsistente Schätzung des ACE, dem Effekt τ in (49) entsprechend, wäre ausgehend von der Formulierung des linearen Regressionsmodell (49) mit Beobachtung der Variablen ξ und entsprechender Konstanthaltung dieser innerhalb des Modells möglich; jedoch folgt aus dem latenten Variablenproblem unmittelbar, dass die Variable ξ lediglich über die manifesten Indikatorvariablen $\vec{X}$ erfassbar ist, die mit ξ durch das lineare faktorenanalytische Modell

$$\vec{X} = \vec{\mu} + \mathbf{\Lambda}\xi + \vec{\delta} \tag{50}$$

in einer Beziehung stehen.

Eine Schätzung des Effektes τ in einem linearen Regressionsmodell wäre damit nur durch eine Regression von Y_i auf den Selektionsindikator S_i sowie auf die manifesten Variablen $\vec{X}$ möglich, die als entsprechende Indikatoren die latente Variable ξ messfehlerbelastet operationalisieren. Offenkundig ist eine konsistente Schätzung von τ in einem solchen Regressionsmodell nur möglich, wenn ein entsprechendes Messmodell, innerhalb dessen die $\vec{X}$ auf die latente Variable ξ regressiert werden, formuliert wird und dem folgend die Messfehler bei der Operationalisierung der latenten Variablen konstant gehalten werden können.

Eine Möglichkeit für eine solche Schätzung des ACE wäre durch die Spezifikation eines linearen Sturkturgleichungsmodells gegeben, das entsprechende Schätzmodell würde den bisherigen Ausführungen mit Abb. 14 übereinstimmen, der interessierende ACE würde dem Effekt des Pfades von S_i auf Y_i entsprechen.

Es ist jedoch unmittelbar ersichtlich, dass bei der Spezifikation eines SEMs eine Vielzahl von Freiheitsgraden bei der Modellierung entsprechend gerichteter Pfade, Kovarianzen und Varianzen zur Verfügung stehen, so dass entsprechend ein Schätzmodell der interessierenden Effekte

genutzt werden könnte, das nicht mit dem zu Grunde liegenden Modell bzw. datengenerierenden Prozess übereinstimmt. Olsson et al. (2000) weisen empirische Verzerrungen und inkonsistente Schätzungen der Effekte auf, die möglichen Fehlspezifikationen und Annahmeverletzungen folgen.

Von diesem Problem ausgehend wird folgend dem parametrischen Verfahren der linearen Strukturgleichungsmodellierung zur Schätzung des ACE bei Vorliegen einer latenten Störvariablen ein Alternativvorschlag gegenübergestellt, der auf dem propensity score, $e(\vec{x}_i)$, beruht. Dieser Vorschlag basiert auf einem mehrstufigen Vorgehen, bei dem zunächst eine Hauptkomponentenanalyse (oder *PCA*, s. folgender Abschnitt) zur Operationalisierung der latenten Variablen ξ über die Korrelationsmatrix der manifesten Variablen $\vec{X}$ durchgeführt wird, wobei für diese Zwecke das noch vorzustellende Eigenwertkriterium zur Extraktion der nötigen Hauptkomponenten verwendet wird. Der Extraktion der Hauptkomponenten folgend wird der propensity score durch eine logistische Regression über die resultierenden Hauptkomponenten geschätzt (*principal components regression*); die resultierenden propensity scores werden für eine Dezentil-Stratifikation genutzt, um entsprechend in den strata den ACE schätzen zu können und den Störeffekt der Variablen ξ, zumindest teilweise, da eine vollständige Konstanthaltung der Messfehler bei einer PCA nicht möglich ist, durch die resultierenden Hauptkomponenten kontrollieren zu können.

Der linearen Strukturgleichungsmodellierung wird bei vollständiger Modellierung des datengenerierenden Prozesses und bei Einhaltung aller Annahmen das vorgeschlagene Verfahren bei der Schätzung des ACE unterlegen sein, da die latente Variable ξ der PCA folgend nur durch maximale Varianzaufklärung der beobachteten Variablen $\vec{X}$ geschätzt und nicht vollständig wie im SEM durch lineare Modelle und Hinzunahme von Messfehler operationalisiert wird. Es sollen folgend jedoch Szenarien simuliert werden, in denen das SEM zur Schätzung des Effektes τ fehlspezifiziert ist und geprüft werden, ob die Schätzung des ACE in den fehlspezifizierten SEMs eine größere Verzerrung aufweist als nach Adjustierung mit dem propensity score, der ausgehend von den Hauptkomponenten geschätzt wird.

5.1.4 Exkurs: Hauptkomponentenanalyse (PCA)

Mit Beobachtung der zentrierten[6] Variablen $\vec{X}^t = (X_1, \ldots, X_p)$ sind die p zugehörigen Hauptkomponenten als diejenigen unkorrelierten Linearkombinationen $F_1, \ldots, F_p$ definiert, die jeweilig sukzessiv maximal Varianz der Variablen $\vec{X}$ erklären. Als die erste Hauptkomponente der Variablen $\vec{X}$, F_1, ist die damit Linearkombination

$$F_1 = a_{11} \cdot X_1 + a_{12} \cdot X_2 + \ldots + a_{1p} \cdot X_p = \vec{a}_1{}^t \vec{X}$$

definiert, deren Varianz, $Var(F_1) = \lambda_1$, im Verhältnis zu allen anderen Hauptkomponenten maximal, die zu den anderen $(p-1)-$Hauptkomponenten orthogonal und deren Koeffizientenvektor normiert ist, so dass $\vec{a}_1{}^t \vec{a}_1 = 1$.

Bereits Pearson (1901) war die Bedeutung der Hauptkomponenten in geometrischer Hinsicht bekannt, jedoch gelang erst Hotelling (1933) die mathematische Definition der Hauptkomponentenanalyse: Da die $j-$te Hauptkomponente, F_j, eine Linearkombination der Variablen $\vec{X}$ ist, steht die Varianz dieser in einem eindeutigen Verhältnis zur Varianz-Kovarianzmatrix der $p-$beobachteten Variablen $\vec{X}$, $\mathbf{\Sigma}_{p \times p}$, es gilt:

$$\begin{aligned} Var(F_j) = \lambda_j &= \vec{a}_j{}^t \mathbf{\Sigma} \vec{a}_j \\ &= a_{j1}^2 \cdot Var(X_1) + a_{j2}^2 \cdot Var(X_2) + \ldots + 2 \cdot a_{j1} \cdot a_{j2} \cdot Cov(X_1, X_2) + \ldots \\ &+ 2 \cdot a_{j(p-1)} \cdot a_{jp} \cdot Cov(X_{p-1}, X_p) \end{aligned}$$

Durch die Anwendung der Methode der *Lagrange*-Multiplikatoren folgt aus den Restriktionen,

$$\lambda_1 > \lambda_2 > \ldots > \lambda_p, \qquad \vec{a}_j^t \vec{a}_j = 1, \qquad Cov(F_j, F_{j'}) = 0, \quad \forall j, j', \text{mit } j \neq j',$$

dass die unbekannten Koeffizienten $\vec{a}_j$ der $j-$ten Hauptkompente durch den $j-$ten normierten Eigenvektor der Matrix $\mathbf{\Sigma}$, $\vec{e}_j$, festgelegt ist, die Varianz λ_j durch den $j-$ten größten Eigenwert von $\mathbf{\Sigma}$. Damit ist die $j-$te Hauptkomponente definiert als

$$F_j = \vec{e}_j{}^t \vec{X} \qquad j = 1, \ldots, p$$

bzw. allgemeingültig:

$$\vec{F} = \mathbf{P}^t \vec{X}, \tag{51}$$

[6] Eine Variable X gilt als zentriert, wenn sie in der Form $X - \mu$ vorliegt, so dass $E(X) = 0$

wobei $\mathbf{P} = [\vec{e}_1, \vec{e}_2, \ldots, \vec{e}_p]$. Es folgt unter Gültigkeit der aufgeführten Restriktionen:

$$\mathrm{tr}(\mathbf{\Sigma}) = \mathrm{tr}(\mathbf{P\Lambda P}^t) = \mathrm{tr}(\mathbf{\Lambda}) = \sum_{j=1}^{p} \lambda_j,$$

wobei $\mathbf{\Lambda}$: Diagonalmatrix der Eigenwerte.

Dem ersichtlich klären die $p-$Hauptkomponenten der Matrix $\mathbf{\Sigma}$ die gesamte Varianz der Variablen $\vec{X}$ auf; dementsprechend lässt sich durch

$$\frac{\lambda_j}{\sum_{j=1}^{p} \lambda_j} \tag{52}$$

der Varianzanteil angeben, der durch die Hauptkomponente F_j erklärt wird.

In geometrischer Hinsicht, wie Pearson (1901) bekannt, entspricht die Ermittlung der Hauptkomponenten einer Rotation des ursprünglichen Koordinatensystems in die Punktewolke der Daten, so dass die erste Hauptkomponente, F_1, den *best fit* an die korrelierenden Daten darstellt. Abbildung 16 verdeutlicht die geometrische Bedeutung im zweidimensionalen Raum.

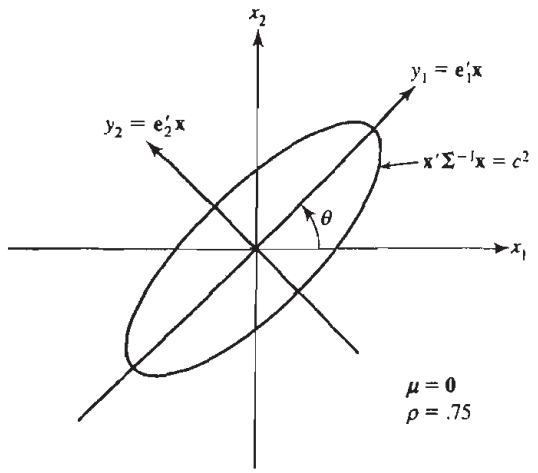

Abbildung 16: Darstellung der Hauptkomponenten im zweidimensionalen Raum, Graphik entnommen aus: Johnson und Wichern (2007)

Mit Durchführung einer Hauptkomponentenanalyse wird zumeist eine Dimensionsreduktion der ursprünglichen Variablen $\vec{X}$ angestrebt, die sich hinsichtlich der Varianzaufklärung (s. (52)) ergibt: Zumeist werden nach Bestimmung der $p-$Hauptkomponenten die ersten $m-$Hauptkomponenten herangezogen, die einen großen Teil der ursprünglichen Varianz der Variablen $\vec{X}$ erklären. In diesem Zusammenhang lässt sich das *Kaiser-Guttman-Kriterium* bzw. *Eigenwertkriterium* einführen, das eine Möglichkeit darstellt, die Anzahl an zu extrahierenden Hauptkomponenten zu bestimmen und damit eine Dimensionsreduktion zu ermöglichen: Insofern die Hauptkomponenten ausgehend von der Korrelationsmatrix $\boldsymbol{\rho}_{p\times p}$ bestimmt werden, lässt sich eine Dimensionsreduktion dahin gehend erreichen, als dass diejenigen Hauptkomponenten entnommen werden, für die $\lambda_j > 1$ (Guttman, 1954).

Unter Rücksichtnahme des Eigenwertkriterium lässt sich $\vec{F}$ partitionieren zu $\vec{F}^t = (\vec{F}_1, \vec{F}_2)$, wobei $\vec{F}_1$ die $m-$Hauptkomponenten mit $\lambda_j > 1$ enthält, $\mathbf{P}$ entsprechend zu $\mathbf{P} = (\mathbf{P}_1, \mathbf{P}_2)$, wobei $\mathbf{P}_1$ als $p \times m-$Matrix die m zugehörigen Eigenvektoren enthält und $\mathbf{\Lambda} = (\mathbf{\Lambda}_1, \mathbf{\Lambda}_2)$, wobei $\mathbf{\Lambda}_1$ als $m \times m-$Diagonalmatrix, die m zugehörigen Eigenwerte enthält. Damit lassen sich die beobachteten Variablen durch die $m-$Hauptkomponenten, die eine Varianz größer als 1 besitzen, darstellen durch das Modell

$$\vec{X} = \vec{\mu} + \mathbf{P}_1 \mathbf{\Lambda}_1^{1/2} \mathbf{\Lambda}_1^{-1/2} \vec{F}_1 + \vec{u}. \tag{53}$$

(53) gleicht damit algebraisch dem allgemeinen faktorenanalytischen Modell, das in (50) definiert worden ist; $\mathbf{P}_1 \mathbf{\Lambda}_1^{1/2}$ entspricht der Ladungsmatrix der Hauptkomponenten (vgl., Yanai & Ichikawa, 2007).

Der Stellenwert der Hauptkomponentenanalyse innerhalb des latenten Variablenproblems ist nicht unumstritten (vgl. hierzu Bollen, 2002); es lässt sich jedoch zeigen, dass die resultierenden Ladungsmatrizen einer Hauptkomponentenanalyse und einer explorativen Faktorenanalyse mit $p \to \infty$ identisch sind (Yanai & Ichikawa, 2007). Johnson und Wichern (2007) verweisen in dem Kapitel zur Schätzung der unbekannten Faktorladungen bei Durchführung einer explorativen Faktorenanalyse auf die Tatsache, dass die Hauptkomponentenanalyse eine Möglichkeit darstellt, initiale Schätzungen für die unbekannten Faktorladungen zu ermöglichen.

Der grundlegende Unterschied, der sich im Verhältnis zur Strukturgleichungsmodellierung bei Durchführung einer Hauptkomponentenanalyse einstellt, ist die fehlende Modellierung von Annahmen über jedwede Fehlerterme: während in Strukturgleichungsmodellen der Fehler, sei es in Form eines Mess- oder Regressionsfehler, als latente Variable in das Schätzmodell aufgenommen und durch entsprechende Skalierungs- und Verteilungsannahmen modelliert wird, verzichtet die Hauptkomponentenanalyse grundlegend auf Annahmen über mögliche Messfehler.

5.2 Schätzung des ACE und datengenerierender Prozess der Simulationen

Im Folgenden wird eine latente Variable ξ als Störvariable definiert, die einerseits mit dem treatment-assignment durch das logistische Modell

$$\Pr(S_i = 1 \mid \xi_i) = \frac{1}{1 + \exp(-(\gamma_0 + \gamma_1 \cdot \xi_i))}, \qquad \text{mit: } \gamma_0 = 0, \gamma_1 = \ln(3),$$

andererseits mit der response-Variablen, Y_i, durch das lineare Modell

$$Y_i = \alpha + \tau \cdot S_i + \gamma_2 \cdot \xi_i + \zeta_i, \qquad \tau = 2, \gamma_2 = 10, \alpha = 0$$

in einer Beziehung steht. τ ist in allen folgenden Simulationsszenarien festgelegt auf $\tau = 2$ und stellt den interessierenden treatment-Effekt dar, der in den Simulationen geschätzt werden soll.

Die Variablen $\vec{X}^t = (X_1, \ldots, X_8)$ dienen im Folgenden als Indikatorvariablen für die latente Variable ξ und werden durch das lineare faktorenanalytische Modell

$$\vec{X} = \vec{\mu} + \mathbf{\Lambda}\xi + \vec{\delta}, \qquad \text{mit: } \vec{\mu} = \vec{0}$$

mit ξ in eine Beziehung gebracht; die Faktorladungsmatrix $\mathbf{\Lambda}$ ist, wenn nicht anderweitig aufgeführt, für alle folgenden Simulationen festgelegt durch

$$\mathbf{\Lambda}^t = \begin{pmatrix} 1.8 & 1.5 & 1.6 & 1.5 & 1.2 & 1.5 & 1.4 & 1.3 \end{pmatrix}.$$

Wenn nicht anderweitig aufgeführt, werden folgende Verteilungsannahmen für die Simulationen festgelegt (Abweichungen hiervon werden jeweilig zu Beginn eines Szenarios spezifiziert):

$$\begin{aligned} &\xi \sim \mathcal{N}(0,1) \\ &\vec{\delta} \sim \mathcal{N}(\vec{0}, \mathbf{\Theta}_\delta^2), \quad \mathbf{\Theta}_\delta^2 = \mathbf{I} \\ &\zeta_i \sim \mathcal{N}(0,1) \end{aligned}$$

Abb. 14 verdeutlicht exemplarisch den datengenerierenden Prozess, es liegen jedoch acht statt drei Indikatorvariablen für die Variable ξ vor. Für die Simulationen werden Stichproben der Umfänge $n = (200, 500, 1000)$ aus den definierten Variablen gezogen (*random-effects*-Modell) und für jeden Stichprobenumfang wird die entsprechende Simulation $k = 500$mal wiederholt. Pro Stichprobe werden folgende Punktschätzungen für den ACE vorgenommen:

Zunächst wird der ACE unadjustiert durch die Differenz der Stichprobenmittel, d.h. $\hat{\tau}_{\text{unadj}} = \bar{Y}_t - \bar{Y}_c$, die Varianz der Mittelwertsdifferenz durch

$$\hat{\sigma}^2_{\bar{Y}_t - \bar{Y}_c} = \left(\frac{1}{n_t} + \frac{1}{n_c}\right) \cdot \frac{(n_t - 1)S_t^2 + (n_c - 1)S_c^2}{n_t + n_c - 2}$$

geschätzt. Die Abweichung des empirischen Mittelwertes der unadjustierten Punktschätzungen, folgend $m(\hat{\tau})$, vom ACE bzw. τ dient folgend als Evidenz für denjenigen bias, der der Konfundierung durch die latente Störvariable resultiert.

Der unadjustierten Schätzung werden zwei unterschiedliche Adjustierungen gegenübergestellt, die hinsichtlich ihrer Güte bei der Schätzung des ACE evaluiert werden sollen:

Zum Einen wird die parametrische Strukturgleichungsmodellierung geprüft, indem in jeder Stichprobe ein Schätzmodell spezifiziert wird, das einerseits aus einem Messmodell zur Operationalisierung der latenten Variablen ξ, andererseits aus einem Strukturmodell, innerhalb dessen die (gerichteten) Effekte der Variablen untereinander formuliert werden, besteht; von Interesse ist der Effekt τ, der im Strukturmodell dem gerichteten Effekt von S_i auf Y_i entspricht. Das jeweilige Strukturgleichungsmodell, das zur Schätzung der Effekte genutzt wird, wird zu Beginn einer Simulation aufgeführt.

Zum Anderen wird eine adjustierte Schätzung des ACE durch eine propensity score-Stratifikation vorgenommen. Hierzu wird im ersten Schritt über die Stichproben-Korrelationsmatrix $\hat{\boldsymbol{\rho}}$ der Indikatorvariablen $\vec{X}$ eine Hauptkomponentenanalyse durchgeführt und diejenigen Hauptkomponenten mit einem Eigenwert $\lambda > 1$ extrahiert.

Die extrahierten Hauptkomponenten, d.h. $F_{j:\lambda_j>1}$, werden im Folgenden in einem logistischen Regressionsmodell als Prädiktorvariablen aufgenommen, um auf diese den treatment-Status der Einheit i regressieren zu können (sog. *principal components regression* (z.B., Massy, 1965)). Es wird damit ersucht, anhand der Hauptkomponenten der $\vec{X}$ die Effekte zur Variablen ξ auf den Selektionsindikator S_i approximieren zu können.

Die resultierenden geschätzten Selektionswahrscheinlichkeiten werden dem folgend als geschätzte propensity scores, $\hat{e}(\vec{x}_i)$, und für eine Dezentil-Stratifikation gegeben dieser propensity scores genutzt: es werden 10 strata basierend auf den geschätzten propensity scores einer Stichprobe definiert, die mit den Dezentilen der propensity scores übereinstimmen. Innerhalb eines stratums wird der stratumsspezifische ACE sowie die zugehörige Varianz der Mittelwertsdifferenz analog zum unadjustierten Fall (s.o.) geschätzt (vgl. hierzu auch (38) sowie (40)). Durch eine gewogene Mittelung dieser schichtspezifischen Schätzungen wird der ACE sowie die Varianz der Mittelwertsdifferenz, folgend für die Ermittlung von Konfidenzintervalle benötigt, stratumsunabhängig geschätzt.

Die Ergebnisse werden folgend tabellarisch und, wenn aussagekräftig, graphisch in Form von 95%igen Stichproben-Konfidenzintervallen berichtet.

Folgende Kennwerte werden in den Tabellen zur Beurteilung der Schätzungen angegeben:

1. Durchschnitt der Punktschätzungen, folgend: $m(\hat{\tau})$

2. empirischer bias, d.h. $m(\hat{\tau}) - \tau$,

3. die prozentuale bias-Reduktion, folgend definiert als

$$\text{bias-Reduktion} = \frac{\text{bias}_{\text{adjustiert}} - \text{bias}_{\text{unadjustiert}}}{\text{bias}_{\text{unadjustiert}}}$$

4. geschätzter M.S.E., wobei

$$\text{M.S.E.} = \text{bias}^2 + \text{Var}(\hat{\tau})$$

5. empirische Standardabweichung der Punktschätzungen als geschätzter Standardfehler,

6. empirische coverage (prozentualer Anteil der 95%igen Stichproben-Konfidenzintervalle, die den Effekt τ überdecken).

Zusätzlich werden in der dritten Simulation die durchschnittlichen geschätzten Faktorladungen der acht Indikatorvariablen berichtet, um mögliche weitere Konsequenzen der Fehlspezifikationen im Messmodell des spezifizierten SEMs aufweisen zu können.

5.3 Ergebnisse

5.3.1 Simulation 1

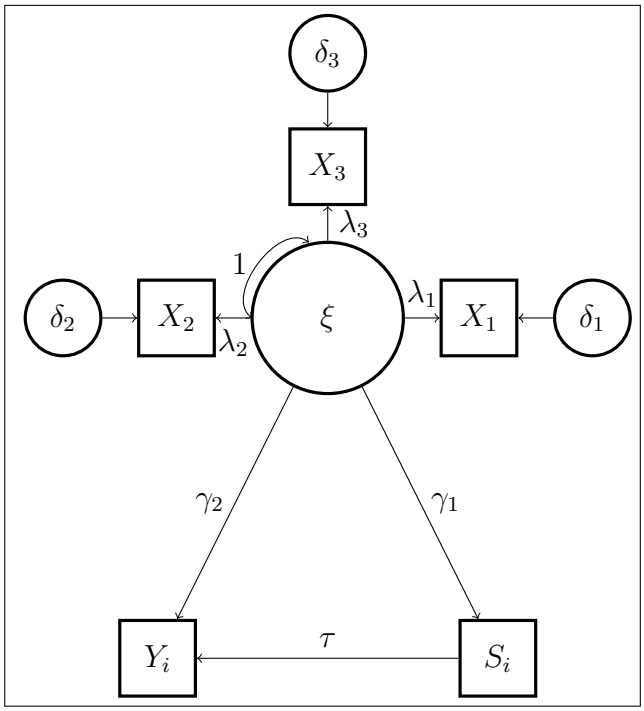

Abbildung 17: SEM der Simulation 1 zur Schätzung der Effekte

In folgender Simulation wird der in Abschnitt 5.2 beschriebene datengenerierende Prozess mit den aufgeführten Verteilungsmodellen zu Grunde gelegt. Das Strukturgleichungsmodell, das als Modell in den Stichproben zur parametrischen Schätzung der Effekte verwendet wird, bildet den datengenerierenden Prozess ab und enthält keine Fehlspezifikationen. Abb. 17 veranschaulicht das verwendete Schätzmodell; die Varianz der latenten Variable ξ wurde in dem Schätzmodell auf 1 fixiert (dem doppelköpfigen Pfeil entsprechend) und es liegen insgesamt acht statt drei Indikatorvariablen zur Operationalisierung der latenten Variablen vor, die in der Graphik der Übersicht halber in der Gänze nicht mit aufgenommen worden sind.

Tabelle 14 stellt die Ergebnisse der Schätzungen des Effektes τ, einerseits für die unadjustierte Schätzung, andererseits nach Adjustierung mit dem SEM und dem vorgeschlagenen propensity score, gerundet auf zwei Nachkommastellen dar. Diese weisen eindeutig auf, dass eine auffällige Reduktion desjenigen bias, der auf die Konfundierung der latenten Störvariablen ξ zurückzuführen ist (vgl. unadjustierte Schätzung), resultiert, insofern der treatment-Effekt adjustiert in dem aufgewiesenen Strukturgleichungsmodell als auch gegeben der vorgeschlagenen propensity score-Stratifikation geschätzt wird.

Es lässt sich konkreter festhalten, dass sich bereits in kleinen Stichproben durchschnittliche Punktschätzungen nach der Adjustierung innerhalb des SEMs, das den datengenerierenden Prozess abbildet, beobachten lassen, die von dem zu schätzenden Effekt τ lediglich auf wenige Nachkommastellen abweichen und damit vernachlässigbare empirische Verzerrungen aufweisen (s. $m(\hat{\tau})$ bzw. bias). Es zeigt sich für alle Stichprobengrößen im Verhältnis zu

den unadjustierten wie auch zu den adjustierten Schätzungen mit dem propensity score, dass die Punktschätzungen des SEMs stets mit den kleinsten geschätzten Standardfehlern einhergehen und damit die verhältnismäßig geringste Variabilität bei der Schätzung des Effektes τ aufweisen. Mit zunehmenden Stichprobenumfang lässt sich eine Abnahme des Standardfehler festhalten; gegeben vernachlässigbarer empirischer Verzerrungen lässt sich mit Abnahme des Standardfehlers eine entsprechende Abnahme der M.S.E.-Schätzungen für eine zunehmende Stichprobengröße beobachten. Für alle Stichprobenumfänge liegen die Überdeckungsraten der geschätzten Konfidenzintervalle in einem mit dem festgelegten Vertrauensniveau von 95% vereinbaren Bereich.

Tabelle 14: Ergebnisse der Simulation 1

	$m(\hat{\tau})$	bias	Red.	S.E.	M.S.E.	cov.
			$n = 200$			
unadjustiert	10.82	8.82	-	1.27	79.41	0%
SEM	2.00	0.00	99.98%	0.40	0.16	96.2%
Stratifikation	2.93	0.93	89.51%	0.51	1.12	62.8%
			$n = 500$			
unadjustiert	10.89	8.89	-	0.74	79.58	0%
SEM	2.00	0.00	100.05%	0.25	0.06	96.6%
Stratifikation	2.95	0.95	89.35%	0.34	1.02	26.2%
			$n = 1000$			
unadjustiert	10.81	8.81	-	0.57	77.94	0%
SEM	2.00	0.00	99.99%	0.17	0.03	97%
Stratifikation	2.95	0.95	89.27%	0.23	0.96	3%

Anm.: $\tau = 2$, $m(\hat{\tau})$: Durchschnitt der Punktschätzungen, Red.: prozentuale bias-Reduktion, S.E.: geschätzter Standardfehler, cov.: prozentuale coverage

Auch für die vorgeschlagene propensity score-Stratifikation gegeben der geschätzten Hauptkomponenten lässt sich eine auffällige Reduktion des bias, der auf die Konfundierung durch

die latente Störvariable ξ zurückzuführen ist, festhalten; die bias-Reduktionen liegen allerdings mit einer Reduktion von ca. 89% eindeutig unterhalb der resultierenden bias-Reduktionen, die der Adjustierung in dem Strukturgleichungsmodell folgen. Im Verhältnis zu den Ergebnissen des SEMs resultieren gegeben der propensity score-Stratifikation für alle Stichprobenumfänge Verzerrungen bei der Schätzung des Effekts τ, die auch mit zunehmenden Stichprobenumfang nicht abnehmen und eine Überschätzung des Effektes τ gegeben dieser Adjustierung aufweisen.

Es lässt sich mit zunehmenden Stichprobenumfang eine einhergehende Abnahme des geschätzten Standardfehlers der mit dem propensity score adjustierten Punktschätzungen berichten. Es resultieren für alle Stichprobenumfänge eindeutige Abweichungen der Überdeckungraten der geschätzten Konfidenzintervalle von dem festgelegten Vertrauensniveau, die in Folge der mit zunehmender Stichprobengröße abnehmenden Standardfehlern und bestehenden empirischen Verzerrungen abnehmen: Bereits in kleinen Stichproben liegt die Überdeckungrate mit ca. 63% unterhalb des festgelegten Vertrauensniveau von 95%; mit $n = 1000$ folgt der bestehenden empirischen Verzerrung und dem verhältnismäßig kleinem Standardfehler, dem entsprechend schmale Schätzintervalle folgen, eine coverage von nur 3%. Abb. 18 verdeutlicht den Befund graphisch durch die Darstellung der Schätzintervalle für $n = 200$ und $n = 1000$:

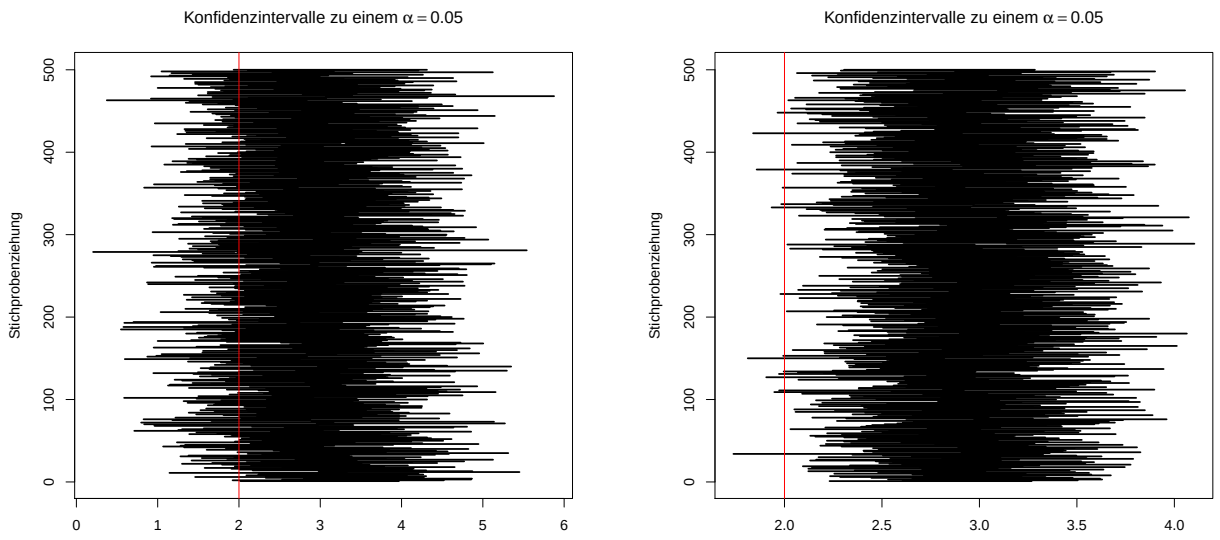

Abbildung 18: 95%ige Stichproben-Konfidenzintervalle der mit dem propensity score adjustierten Schätzungen für $n = 200$ (links) und $n = 1000$ (rechts)

5.3.2 Simulation 2

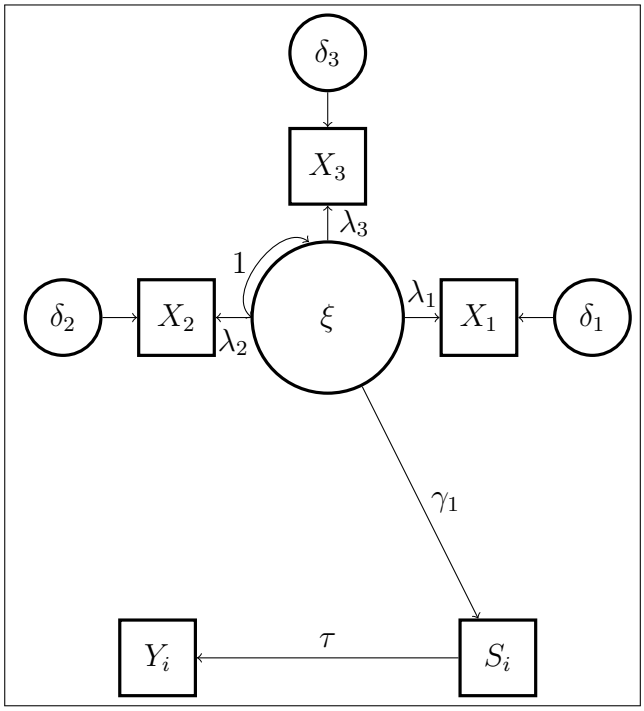

Abbildung 19: SEM der Simulation 2 zur Schätzung der Effekte

In folgender Simulation wird der in Abschnitt 5.2 beschriebene datengenerierende Prozess beibehalten. Im Vergleich zur ersten Simulation wird folgend das Strukturgleichungsmodell, das zur parametrischen Schätzung des Effektes τ genutzt wird, ohne den Effekt γ_2 spezifiziert (vgl. Abb. 17, in der der Effekt weiterhin aufgeführt wird). Damit wird folgend eine mögliche Fehlspezifikation des Strukturgleichungsmodells hinsichtlich des datengenerierenden Prozesses simuliert, die der Auslassung einer der entscheidenden Pfade - dem Effekt der Variablen ξ auf die response-Variable Y_i - im Strukturmodell entspricht. Abb. 19 veranschaulicht das SEM, das in den Stichproben zur Adjustierung spezifiziert wird.

Tabelle 15 führt die Ergebnisse der Schätzungen des Effektes τ, gerundet auf zwei Nachkommastellen, für den unadjustierten Fall sowie für die Adjustierungen im SEM als auch gegeben der vorgeschlagenen propensity score-Stratifikation auf.

Die Ergebnisse weisen unmittelbar die weitreichenden Konsequenzen des ausgelassenen Pfades bei der Spezifikation des SEMs auf: Die durchschnittlichen adjustierten Punktschätzungen des interessierenden Effektes τ im SEM decken sich mit den durchschnittlichen unadjustierten Schätzungen. Damit ergeben sich für alle Stichprobengrößen empirische Verzerrungen gegeben der parametrischen Adjustierung im SEM, die identisch sind mit denen der unadjustierten Schätzung und somit aufweisen, dass die von dem datengenerierenden Prozess abweichende Spezifikation des SEMs hinsichtlich einer möglichen Reduktion des bias, der auf die konfundierende Störvariable zurückgeführt werden kann, keinen begünstigenden Effekt besitzt. Es lässt sich zusätzlich zu den aufgewiesenen Verzerrungen festhalten, dass sämtliche anderen Kennwerte, die zur Beurteilung der Schätzungen herangezogen werden, deckungsgleich mit denen der

unadjustierten Schätzung des Effektes sind: So lässt sich festhalten, dass für alle Stichprobengrößen die resultierenden Standardfehler, M.S.E.-Schätzungen sowie Überdeckungsraten denen der unadjustierten Schätzung gleichen.

Tabelle 15: Ergebnisse der Simulation 2

	$m(\hat{\tau})$	bias	Red.	S.E.	M.S.E.	cov.
			$n = 200$			
unadjustiert	10.82	8.82	-	1.27	79.42	0%
SEM	10.82	8.82	0.00%	1.27	79.42	0%
Stratifikation	2.93	0.93	89.51%	0.51	1.11	62.8%
			$n = 500$			
unadjustiert	10.89	8.89	-	0.74	79.52	0%
SEM	10.89	8.89	0.00%	0.74	79.52	0%
Stratifikation	2.95	0.95	89.35%	0.34	1.01	26.2%
			$n = 1000$			
unadjustiert	10.81	8.81	-	0.57	78.01	0%
SEM	10.81	8.81	0.00%	0.57	78.01	0%
Stratifikation	2.95	0.95	89.27%	0.23	0.95	3%

Anm.: $\tau = 2$, $m(\hat{\tau})$: Durchschnitt der Punktschätzungen, Red.: prozentuale bias-Reduktion, S.E.: geschätzter Standardfehler, cov.: prozentuale coverage

Dem gegenüber stehen die Ergebnisse der adjustierten Schätzung des Effektes τ gegeben der vorgeschlagenen propensity score-Stratifikation: Es zeigt sich in Anbetracht der Modellierung des propensity scores ausgehend von den resultierenden Hauptkomponenten, dass eine auffällige Reduktion der Konfundierung durch die latente Störvariable ξ möglich ist. Bereits in kleinen Stichproben resultiert eine bias-Reduktion von ca. 89%, die einen entsprechend begünstigenden Effekt hinsichtlich einer adjustierten Schätzung gegeben der propensity score-Stratifikation aufweist. An dieser Stelle ist jedoch eindeutig hervorzuheben, dass für alle Stichprobenumfänge Verzerrungen resultieren, die, gegeben abnehmender Standardfehler und bestehender Verzerrung bei zunehmenden Stichprobenumfang, eine systematische Überschätzung

des interessierenden Effektes aufweisen. In Konsequenz folgen dem Trend bestehender Verzerrungen und abnehmender Standardfehler Intervallschätzungen, die in großen Stichproben mit einer Überdeckungsrate von nur 3% einhergehen.

5.3.3 Simulation 3

In der folgenden Simulation wurde der in Abschnitt 5.2 beschriebene datengenerierende Prozess wie folgt modifiziert: Es wurden zwei latente Variablen, folgend ξ und η, definiert, die durch das lineare Modell

$$\eta = 0.8 \cdot \xi + \zeta_1$$

miteinander in einer Beziehung stehen. Die Indikatorvariablen $\vec{X}^t = (X_1, \ldots, X_5)$ stehen durch das lineare faktorenanalytische Modell

$$\vec{X} = \vec{\mu}_{\vec{X}} + \mathbf{\Lambda}_{\vec{X}} \vec{\xi} + \vec{\delta}$$

mit der latenten Variablen ξ in einer Beziehung, die Indikatorvariablen $\vec{Y}^t = (X_6, \ldots, X_8)$ durch das lineare faktorenanalytische Modell

$$\vec{Y} = \vec{\mu}_{\vec{Y}} + \mathbf{\Lambda}_{\vec{Y}} \vec{\eta} + \vec{\epsilon}$$

mit der latenten Variablen η.

In dem datengenerierenden Prozess wurde die latente Variable η als konfundierende Störvariable spezifiziert; das treatment-assignment wurde in Abhängigkeit von η durch das logistische Modell:

$$\Pr(S_i = 1 \mid \eta_i) = \frac{1}{1 + \exp(-(\gamma_0 + \gamma_1 \cdot \eta_i))} \qquad \text{mit: } \gamma_0 = 0, \gamma_1 = \ln(3),$$

die response-Variable Y_i durch das lineare Modell

$$Y_i = \alpha + \tau \cdot S_i + \gamma_2 \cdot \eta_i + \zeta_2 \qquad \text{wobei: } \alpha = 0, \tau = 2, \gamma_2 = 10$$

realisiert. Folgende Verteilungsmodelle und Modellspezifikationen wurden für den datengenerierenden Prozess gewählt:

$$\xi_i \sim \mathcal{N}(0, 1)$$

$$\vec{\mu}_{\vec{X}} = \vec{\mu}_{\vec{Y}} = \vec{0}$$

$$\vec{\delta} \sim \mathcal{N}(\vec{0}, \mathbf{I}), \quad \vec{\epsilon} \sim \mathcal{N}(0, \mathbf{I}), \quad \vec{\zeta} \sim \mathcal{N}(0, \mathbf{I})$$

$$\mathbf{\Lambda}_{\vec{X}}^{t} = (1.8 \;\; 1.5 \;\; 1.6 \;\; 1.5 \;\; 1.2)\,, \quad \mathbf{\Lambda}_{\vec{Y}}^{t} = (1.5 \;\; 1.4 \;\; 1.3)$$

In den Stichproben wurde das lineare Strukturgleichungsmodell zur parametrischen Schätzung der Effekte dergestalt spezifiziert, dass die latente Variable ξ - anstelle der Variablen η - als konfundierende Störvariable in dem Strukturmodell klassifiziert und - anstelle eines gerichteten Effektes von ξ auf η - eine Korrelation zwischen den beiden latenten Variablen ξ und η formuliert wurde. Abb. 20 veranschaulicht das spezifizierte SEM in den Stichproben, das mit damit eindeutig von dem datengenerierenden Prozess abweicht.

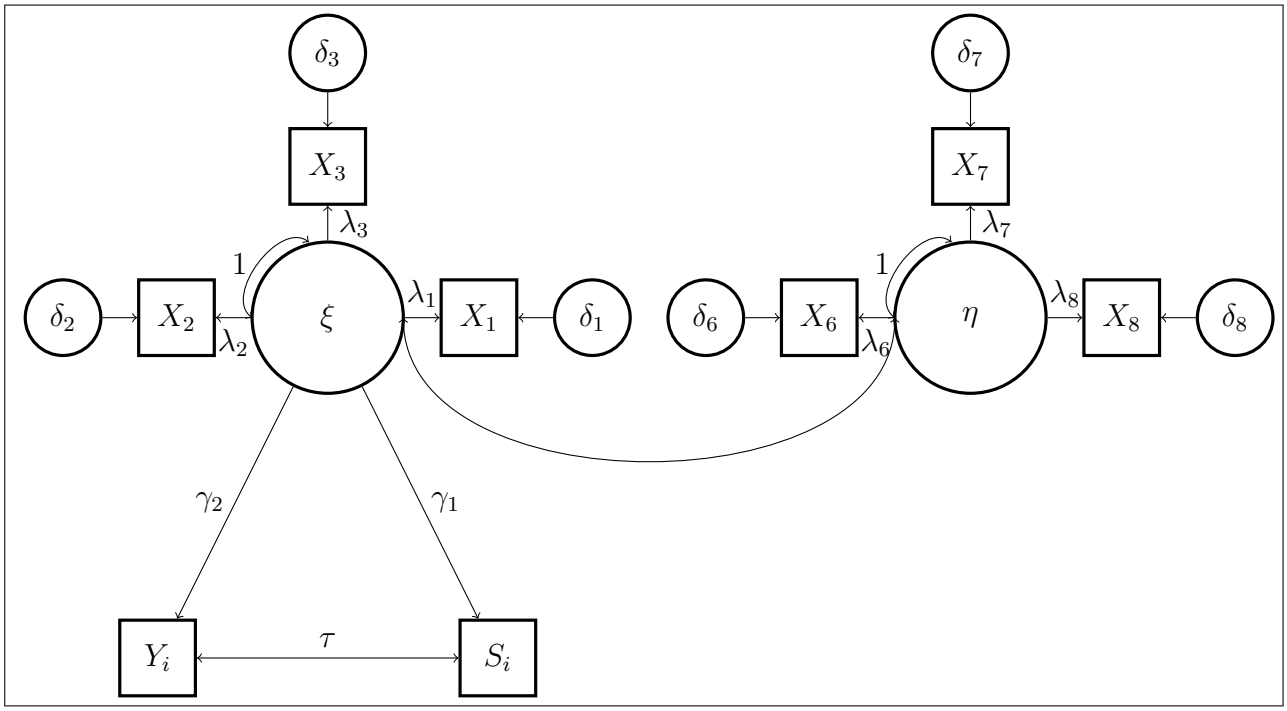

Abbildung 20: SEM der Simulation 3 zur Schätzung der Effekte

Tabelle 16 führt, auf zwei Nachkommastellen gerundet, die Ergebnisse der Schätzungen des Effekts τ, einerseits für den unadjustierten Fall, andererseits für die Adjustierungen im SEM sowie gegeben der propensity score-Stratifikation auf.

Für die adjustierten Punktschätzungen des von dem datengenerierenden Prozess abweichend formulierten SEMs lassen sich auffällige Verzerrungen bei der Schätzung des Effektes τ festhalten: So resultieren durchschnittliche Punktschätzungen nach der Adjustierung im SEM, die in kleinen Stichproben einen auffälligen bias aufweisen, der auch mit zunehmenden Stichprobenumfang nicht abnimmt. Dementsprechend folgen für alle Stichprobengrößen lediglich bias-Reduktionen von ca. 32% gegeben einer Adjustierung im SEM. Diesem Ergebnis gegenüber stehen eindeutig größere bias-Reduktionen von ca. 84% gegeben der propensity score-Stratifikation.

Entgegen der ersten Simulation und im Vergleich zur propensity score-Stratifikation lassen sich im Falle der vorliegenden Fehlspezifikation des SEMs größere Standardfehler festhalten, so dass die mit dem propensity score adjustierten Punktschätzungen für alle Stichprobengrößen die verhältnismäßig geringste Variabilität aufweisen. Es lässt sich die Tendenz abnehmender Standardfehler mit zunehmenden Stichprobenumfang für die geschätzten Effekte des SEMs festhalten.

Tabelle 16: Ergebnisse der Simulation 3

	$m(\hat{\tau})$	bias	Red.	S.E.	M.S.E.	cov.
			$n = 200$			
unadjustiert	15.00	13.00	-	1.56	171.51	0%
SEM	10.80	8.80	32.35%	1.37	79.26	0%
Stratifikation	4.01	2.01	84.57%	0.86	4.76	36.4%
			$n = 500$			
unadjustiert	15.08	13.08	-	0.97	172.16	0%
SEM	10.91	8.91	31.91%	0.91	80.20	0%
Stratifikation	4.04	2.04	84.42%	0.51	4.42	4.2%
			$n = 1000$			
unadjustiert	15.11	13.11	-	0.72	172.48	0%
SEM	10.90	8.90	32.14%	0.61	79.55	0%
Stratifikation	4.06	2.06	84.31%	0.37	4.37	0%

Anm.: $\tau = 2$, $m(\hat{\tau})$: Durchschnitt der Punktschätzungen, Red.: prozentuale bias-Reduktion, S.E.: geschätzter Standardfehler, cov.: prozentuale coverage

Die Ergebnisse der propensity score-Stratifikation weisen einerseits auf, dass eine Reduktion der Konfundierung durch die latente Störvariable η mit dieser Adjustierungsmethode möglich ist, andererseits jedoch merkliche Verzerrungen bei der Schätzung des Effektes τ bestehen bleiben. Zwar liegen diese Verzerrungen bereits in kleinen Stichproben unterhalb der Verzerrungen, die der Adjustierung im SEM folgen, jedoch lässt sich gegeben dieser Adjustierungsmethode auch keine Abnahme der Verzerrungen mit zunehmenden Stichprobenumfang festhalten.

Gegeben bestehender Verzerrungen und mit zunehmenden Stichprobenumfang abnehmenden Standardfehlern folgen Intervallschätzungen, die in Stichproben der Größe $n = 1000$ nicht mehr den zu schätzenden Effekt τ überdecken.

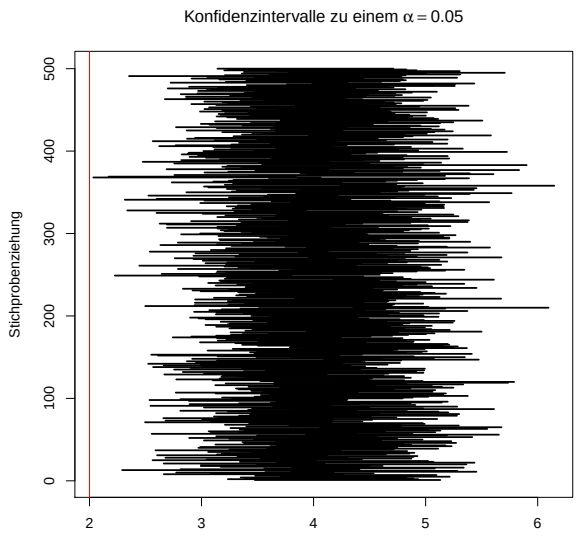

Abbildung 21: Konfidenzintervalle zur Schätzung des Effektes τ nach propensity score-Stratifikation ($n = 1000$)

Damit lassen sich die bestehenden Verzerrungen eindeutig auf eine systematische Überschätzung des Effektes τ zurückführen und nicht auf vereinzelte extreme Schätzungen (vgl. Abb. 21).

Zusätzlich zu den obig aufgewiesenen Verzerrungen bei der Schätzung des Effektes τ, die der vom datengenerierenden Prozess abweichenden Spezifikation des Strukturmodells im SEM folgen, lassen sich Verzerrungen bei der Schätzung einiger Faktorladungen des Messmodells im SEM berichten, die kurz aufgeführt werden sollen (vgl. Tabelle 17). In Anbetracht der geschätzten Faktorladungen zeigt sich, dass die Spezifikation einer Korrelation - anstelle eines gerichteten Pfades - zwischen beiden latenten Variablen des Messmodells zu Verzerrungen der geschätzten Faktorladungen der Indikatorvariablen, die der latenten Variablen η zugeordnet sind, führt (vgl. $m(\hat{\lambda}_6)$, $m(\hat{\lambda}_7)$, $m(\hat{\lambda}_8)$).

Tabelle 17: geschätzte Faktorladungen - Simulation 3

n	$m(\hat{\lambda}_1)$	$m(\hat{\lambda}_2)$	$m(\hat{\lambda}_3)$	$m(\hat{\lambda}_4)$	$m(\hat{\lambda}_5)$	$m(\hat{\lambda}_6)$	$m(\hat{\lambda}_7)$	$m(\hat{\lambda}_8)$
200	1.79	1.50	1.59	1.49	1.20	1.92	1.72	1.66
500	1.81	1.50	1.60	1.50	1.20	1.92	1.79	1.66
1000	1.80	1.51	1.60	1.50	1.20	1.92	1.80	1.67

5.3.4 Simulation 4

In der folgenden Simulation wird der in Abschnitt 5.2 beschriebene datengenerierende Prozess modifiziert, so dass das von der latenten Variablen ξ abhängige treatment-assignment durch das logistische Modell

$$\Pr(S_i = 1 \mid \xi_i) = \frac{1}{1 + \exp(-(\gamma_0 + \gamma_1 \cdot \xi_i^2))} \qquad \text{mit: } \gamma_0 = 0, \gamma_1 = \ln(3),$$

das Modell der response-Variablen Y_i durch das Modell

$$Y_i = \alpha + \tau \cdot S_i + \gamma_2 \cdot \xi_i^2 + \zeta_i \qquad \text{mit: } \alpha = 0, \tau = 2, \gamma_2 = 10$$

realisiert wird. Dem ersichtlich werden in dem datengenerierenden Prozess die einfachen linearen Effekte, die in den vorherigen Simulationen simuliert worden sind, durch Modelle mit einem quadratischen Effekt der Variablen ξ ersetzt. Sämtliche anderen Spezifikationen, die in Abschnitt 5.2 aufgeführt worden sind, werden beibehalten.

Das lineare Strukturgleichungsmodell zur parametrischen Schätzung des Effektes τ wird wie in Simulation 1 (vgl. Abb. 17) spezifiziert; es werden somit einfache lineare Effekte der latenten Variablen ξ auf die Variablen Y_i und S_i geschätzt. Tabelle 18 führt die Ergebnisse der unadjustierten sowie der adjustierten Schätzungen des Effekts τ, einerseits nach Adjustierung im SEM, andererseits nach Adjustierung mit dem vorgeschlagenen propensity score, gerundet auf zwei Nachkommastellen auf.

Die resultierenden Verzerrungen, die sich nach der Adjustierung im SEM bei der Schätzung des Effektes τ ergeben, weisen eindeutig auf, dass die Aufnahme quadratischer Effekte in dem datengenerierenden Prozess, insofern diese nicht bei der Spezifikation des SEMs in den Stichproben berücksichtigt werden, weitreichende Konsequenzen hat: Die durchschnittlichen Punktschätzungen weichen von den Ergebnissen der unadjustierten Schätzung nur auf wenige Nachkommastellen ab und führen damit auf, dass eine Adjustierung im SEM keinen nennenswerten begünstigenden Effekt hinsichtlich einer möglichen Reduktion des bias, der sich auf den konfundierenden Effekt der latenten Störvariable zurückführen lässt, hat. Es lässt sich festhalten, dass die resultierenden bias-Reduktionen mit zunehmenden Stichprobenumfang abnehmen, so dass lediglich eine Reduktion von ca. 0.3% des bias, der der Konfundierung der latenten Störvariablen resultiert, der Adjustierung im SEM in Stichproben der Größe $n = 1000$ folgt. Des weiteren führt die ausgelassene Spezifikation der quadratischen Effekte in dem SEM zu adjustierten Punktschätzungen, die in Standardfehlern resultieren, die mit denen der unadjustierten

Punktschätzungen des Effektes übereinstimmen. In Folge der auffälligen Verzerrungen resultieren Intervallschätzungen, die gegeben aller Stichprobengrößen mit einer Überdeckungsrate von 0% einhergehen und damit eine systematische Überschätzung des Effektes τ aufweisen (vgl. Abb. 22 für die Darstellung der Schätzintervalle der Stichproben vom Umfang $n = 1000$). Im Verhältnis zur ersten Simulation, in der das SEM in den Stichproben den datengenerierenden Prozess vollständig abbildete, weisen die vorliegenden Ergebnisse keine empirische M.S.E.-Konvergenz auf: Mit $n = 1000$ resultiert gegeben der bestehenden Verzerrung ein geschätzter M.S.E. von 89.93; dieser ist damit geringfügig größer als die resultierende M.S.E.-Schätzung bei $n = 200$.

Tabelle 18: Ergebnisse der Simulation 4

	$m(\hat{\tau})$	bias	Red.	S.E.	M.S.E.	cov.
			$n = 200$			
unadjustiert	11.14	9.14	-	1.59	86.05	0%
SEM	11.02	9.02	1.31%	1.60	83.89	0%
Stratifikation	4.35	2.35	74.33%	1.23	7.02	55.8%
			$n = 500$			
unadjustiert	11.09	9.09	-	0.90	83.52	0%
SEM	11.06	9.06	0.42%	0.90	82.84	0%
Stratifikation	5.44	3.44	62.19%	0.92	12.68	19.6%
			$n = 1000$			
unadjustiert	11.16	9.16	-	0.70	84.42	0%
SEM	11.13	9.13	0.29%	0.70	83.93	0%
Stratifikation	5.72	3.72	59.41%	0.55	14.13	0.8%

Anm.: $\tau = 2$, $m(\hat{\tau})$: Durchschnitt der Punktschätzungen, Red.: prozentuale bias-Reduktion, S.E.: geschätzter Standardfehler, cov.: prozentuale coverage

Hinsichtlich einer möglichen Reduktion der Konfundierung durch die latente Störvariable weisen die Ergebnisse der propensity score-Stratifikation im Vergleich zu denen des SEMs eine auffällige bias-Reduktion auf. Bereits in kleinen Stichproben lässt sich eine bias-Reduktion

von ca. 74% berichten, wobei hervorgehoben werden muss, dass sich der propensity score-Stratifikation folgend auffällige Verzerrungen bei der Schätzung des Effektes τ beobachten lassen, die mit zunehmenden Stichprobenumfang zunehmen, jedoch stets unterhalb der Verzerrungen des SEMs liegen. In Folge zunehmender Verzerrungen und abnehmender Standardfehler lassen sich mit zunehmenden Stichprobenumfang zunehmende M.S.E.-Werte berichten, die eine systematische Überschätzung des Effektes τ und eine Zunahme in den resultierenden Verzerrungen aufweisen. Dieser Tendenz entsprechend nehmen die Überdeckungsraten der geschätzten Konfidenzintervalle mit zunehmenden Stichprobenumfang ab und es resultieren in Stichproben vom Umfang $n = 1000$ Schätzintervalle, die nur in 0.8% der Schätzungen den Effekt τ überlagern (vgl. Abb. 22).

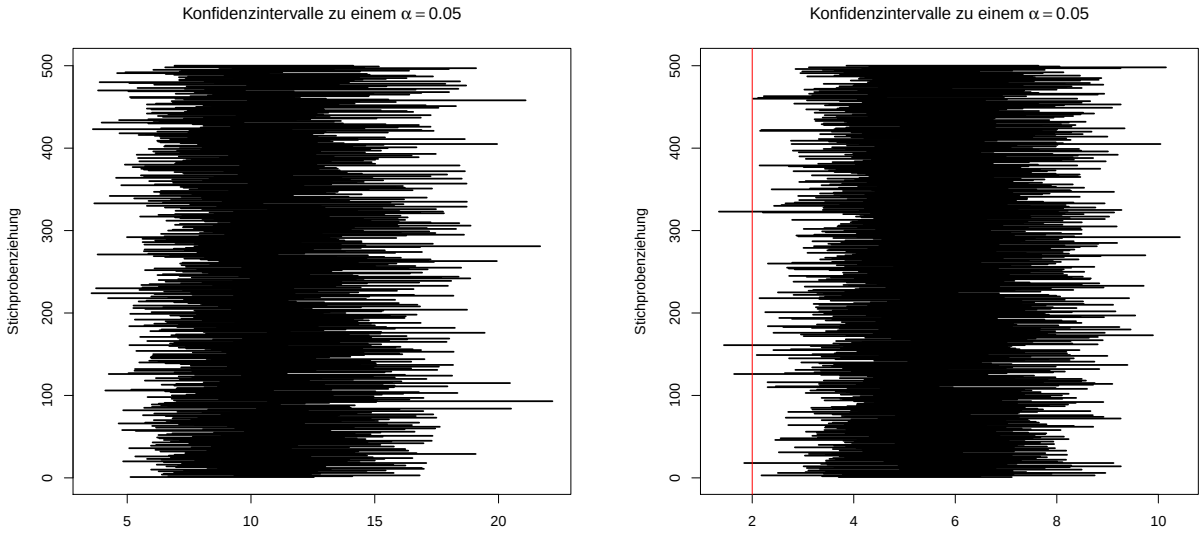

Abbildung 22: Konfidenzintervalle nach Adjustierung im SEM (links) sowie nach propensity score-Stratifikation (rechts)

5.3.5 Simulation 5

Unter Anderem werden an den datengenerierenden Prozess bei der Spezifikation eines Strukturgleichungsmodells die Annahmen der linearen Unabhängigkeit sämtlicher Fehlerterme untereinander, dies beinhaltet sowohl die Mess- als auch die Regressionsfehler, sowie der Unabhängigkeit der Fehlerterme von jeglichen exogenen und endogenen Variablen im Mess- und Strukturmodell des SEMs zu Grunde gelegt. In der folgenden Simulation soll eine Verletzung der aufgeführten Modellannahmen hinsichtlich der Schätzung des Effektes τ untersucht werden; hierzu wird der in Abschnitt 5.2 beschriebene datengenerierende Prozess wie folgt modifiziert: Das Messmodell $\vec{X} = \vec{\mu} + \mathbf{\Lambda}\vec{\xi} + \vec{\delta}$ wird dahin gehend spezifiziert, als dass die Messfehler der Indikatorvariablen X_1 und X_2, δ_1 und δ_2, durch das lineare Modell

$$\delta_2 = 0.6 \cdot \delta_1 + \delta_2^* \qquad \delta_2^* \sim \mathcal{N}(0,1)$$

in einer Beziehung zueinander stehen; das resultierende Messmodell der Indikatorvariablen X_2 lautet entsprechend:

$$X_2 = 1.5 \cdot \xi + (0.6 \cdot \delta_1 + \delta_2^*).$$

Der Fehler δ_2 wird zusätzlich in das logistische Modell des treatment-assignments mit aufgenommen, so dass folgend:

$$\Pr(S_i = 1 \mid \xi_i, \delta_2) = \frac{1}{1 + \exp(-(\gamma_0 + \gamma_1 \cdot \xi_i + \delta_2))} \quad \text{mit: } \gamma_0 = 0, \gamma_1 = \ln(3).$$

Sämtliche andere Spezifikationen und Verteilungsannahmen, die in Abschnitt 5.2 beschrieben worden sind, wurden beibehalten. Das in den Stichproben modellierte SEM zur parametrischen Schätzung des Effektes τ wird wie in der ersten Simulation ohne Rücksicht auf den modifizierten datengenerierenden Prozess spezifiziert (vgl. Abb. 17).

Tabelle 19 führt die Ergebnisse der Schätzungen des Effektes τ, gerundet auf zwei Nachkommastellen, zum Einen für den unadjustierten Fall, zum Anderen nach den Adjustierungen im SEM sowie nach der propensity score-Stratifikation auf. Unmittelbar ersichtlich in Anbetracht der resultierenden negativen Verzerrungen ist eine Unterschätzung des Effektes τ, die der Adjustierung im SEM folgt. Sämtliche bias-Schätzungen liegen unterhalb des Effekts τ und bleiben auch mit zunehmenden Stichprobenumfang konstant. Gegeben der bestehenden Verzerrungen und der, mit zunehmenden Stichprobenumfang abnehmender, Standardfehler lässt sich die aufgewiesene Unterschätzung als systematisch einordnen; ein Befund, der durch die abnehmende coverage der geschätzten Konfidenzintervalle gestützt wird.

Tabelle 19: Ergebnisse der Simulation 5

	$m(\hat{\tau})$	bias	Red.	S.E.	M.S.E.	cov.
			$n = 200$			
unadjustiert	9.56	7.56	-	1.29	58.77	0%
SEM	1.25	-0.75	109.93%	0.45	0.77	57%
Stratifikation	2.04	0.04	99.46%	0.55	0.30	95.8%
			$n = 500$			
unadjustiert	9.59	7.59	-	0.86	58.42	0%
SEM	1.24	-0.76	109.99%	0.27	0.65	15.2%
Stratifikation	2.03	0.03	99.63%	0.32	0.10	97.4%
			$n = 1000$			
unadjustiert	9.51	7.51	-	0.58	56.68	0%
SEM	1.25	-0.75	110.04%	0.19	0.60	1.8%
Stratifikation	2.05	0.05	99.32%	0.23	0.05	95.8%

Anm.: $\tau = 2$, $m(\hat{\tau})$: Durchschnitt der Punktschätzungen, Red.: prozentuale bias-Reduktion, S.E.: geschätzter Standardfehler, cov.: prozentuale coverage

Abb. 23 verdeutlicht die aufgewiesene systematische Unterschätzung des Effektes in den Stichproben. Diesem Ergebnis gegenüber stehen die Ergebnisse der Schätzungen nach der propensity score-Stratifikation: Bereits in kleinen Stichproben weisen die resultierenden Verzerrungen geringfügige Abweichungen von dem zu schätzenden Effekt τ auf, so dass sich in Stichproben vom Umfang $n = 200$ eine Reduktion von ca. 99.5% des bias, der auf die konfundierende Störvariable zurückzuführen ist, festhalten lässt. Für alle Stichprobenumfänge resultieren gegeben der propensity score-Stratifikation Standardfehler, die über den Standardfehlern der im SEM adjustierten Punktschätzungen liegen. Es lässt sich die Tendenz abnehmender Standardfehler mit zunehmenden Stichprobenumfang festhalten; gegeben der geringfügigen Verzerrungen liegen trotz abnehmender Standardfehler sämtliche resultierende Überdeckungsraten in einem mit dem Vertrauensniveau von 95% vereinbaren Bereich.

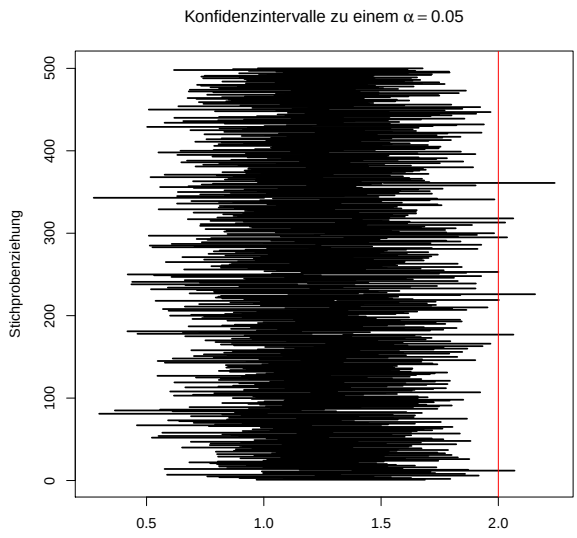

Abbildung 23: Schätzintervalle nach Adjustierung im SEM (n = 1000)

5.4 Diskussion

Unter der Voraussetzung der Spezifikation eines Strukturgleichungsmodells, das mit dem datengenerierenden Prozess übereinstimmt, weisen die Ergebnisse des ersten Simulationsszenarios die Überlegenheit der Strukturgleichungsmodellierung als parametrisches Adjustierungsverfahren bei der Schätzung des Effektes τ im Verhältnis zu der vorgeschlagenen propensity score-Stratifikation auf: Die resultierenden Verzerrungen, die sich bereits in kleinen Stichproben nach der Adjustierung im SEM beobachten lassen, weichen lediglich auf wenige Nachkommastellen von dem zu schätzenden Effekt τ ab und weisen zudem eine effiziente Schätzung des interessierenden Effektes auf, die sich in den verhältnismäßig kleinsten Standardfehlern für alle Stichprobengrößen äußert.

Diesen Ergebnissen gegenüber stehen die der vorschlagenen propensity score-Stratifikation, die eine Möglichkeit zur Schätzung des treatment-Effekts auf Grundlage der Hauptkomponenten und der in Folge damit geschätzten Selektionswahrscheinlichkeiten bietet und auf eine vollständige Modellierung des datengenerierenden Prozesses, wie im SEM notwendig, verzichtet. Bereits in kleinen Stichproben lässt sich eine auffällige Reduktion desjenigen bias, der auf die Konfundierung der latenten Variable zurückgeführt werden kann, festhalten, so dass bedingt das Hauptziel einer Adjustierung bei der Schätzung des treatment-Effektes gewährleistet wird; denn auch in großen Stichproben folgen der propensity score-Stratifikation auffällige Verzer-

rungen, die sich in dreierlei Hinsicht erklären lassen: Zum Einen wird in diesem Vorschlag zur Adjustierung die konfundierende latente Störvariable nicht in einem Messmodell mit Spezifikation jeglicher Messfehler, die bei der Schätzung der von der latenten Variablen ausgehenden Effekte konstant gehalten werden müssen, operationalisiert sondern wird durch die Extraktion der Hauptkomponenten der zugehörigen Indikatorvariablen hinsichtlich einer maximalen Varianzaufklärung approximiert, zum Anderen fließen in Folge dieser Approximation der latenten Variablen nicht alle relevanten Informationen über das treatment-assignment, das durch diese latente Variable bedingt wird, in die Schätzung des propensity scores mit ein. Damit führt die im Verhältnis zu einer Modellierung des datengenerierenden Prozesses innerhalb eines SEMs vereinfachte Schätzung des treatment-Effektes eindeutig zu einem Informationsverlust hinsichtlich des zu Grunde liegenden treatment-assignments, der sich offenkundig in den entsprechenden Verzerrungen bei der Schätzung des treatment-Effekts niederschlägt. Des Weiteren handelt es sich, wie bereits theoretisch aufgewiesen, bei der propensity score-Stratifikation um ein Adjustierungsverfahren, das stets zu inkonsistenten Ergebnissen führt, da innerhalb eines stratums statistische Einheiten mit ähnlichen, aber nicht identischen propensity scores gruppiert werden und somit auch in großen Stichproben geringfügige Konfundierungen der Störvariablen resultieren.

Die Ergebnisse der weiteren Simulationsszenarien weisen jedoch die Konsequenzen auf, die sich in Folge einer von dem datengenerierenden Prozess abweichenden Spezifikation des Strukturgleichungsmodells ergeben und die, verglichen mit diesen, die Ergebnisse der mit dem propensity score adjustierten Schätzungen des treatment-Effektes begünstigend hervorheben. In diesen Szenarien wurden mögliche Fehlspezifikationen der Strukturgleichungsmodellierung simuliert, die sich hinsichtlich der Freiheitsgrade, die bei der Modellierung des datengenerierenden Prozesses in den Stichproben zur Verfügung stehen, ergeben: Eine dieser möglichen Fehlspezifikationen ist das Szenario eines *omitted path*, d.h. der fehlenden Spezifikation eines der relevanten Pfade, die bei der Schätzung des Effektes τ konstant gehalten werden müssen, um konsistente Schätzergebnisse beobachten zu können (Simulation 2). In dieser Simulation wurde das SEM in den Stichproben dergestalt modelliert, dass der Pfad γ_2, der den gerichteten Effekt der latenten Variablen ξ auf die response-Variable Y_i angibt, bei der Spezifikation ausgelassen worden ist und dem folgend bei der Schätzung des Effektes τ nicht konstant gehalten werden konnte. Dieser möglichen Fehlspezifikation resultierten Punktschätzungen, die denen der unadjustierten

Punktschätzung glichen und damit keinen begünstigenden Effekt hinsichtlich einer Reduktion desjenigen bias, der auf die latente Störvariable ξ zurückgeführt werden kann, aufwiesen. Trotz abnehmender Standardfehler resultierten auch in großen Stichproben Verzerrungen, die identisch mit denen der unadjustierten Schätzung waren und somit eine fehlende M.S.E.-Konsistenz, soweit dies in den Simulationsstudien ausgesagt werden kann, aufwiesen. Dem gegenüber standen auffällige bias-Reduktionen, die der vorgeschlagenen propensity score-Stratifikation folgten. Jedoch muss betont werden, dass dieser Vorschlag zur adjustierten Schätzung einerseits mit einer großen bias-Reduktion einhergeht, andererseits die beobachtbaren Verzerrungen auch in großen Stichproben resultierten und damit stets eine Überschätzung des Effektes mit dieser Adjustierung einhergeht.

Während in den datengenerierenden Prozessen der anderen Simulationsszenarien einfache lineare Effekte der latenten Störvariablen ξ auf die response-Variable Y_i sowie auf das treatment-assignment modelliert worden sind, wurde in der vierten Simulation der datengenerierende Prozess mit einem quadratischen Effekt der Variablen ξ auf beide Variablen spezifiziert. Als mögliche Fehlspezifikation wurden in dem SEM weiterhin einfache lineare Effekte geschätzt, so dass die entsprechende Linearitätsannahme, die dieser Schätzung zu Grunde liegt, verletzt wird. Die Ergebnisse dieser Simulation weisen eindeutig die weitreichenden Konsequenzen dieser Annahmeverletzung auf, da sich Verzerrungen bei der Schätzung des treatment-Effektes beobachten lassen, die nur auf wenige Nachkommastellen von denen der unadjustierten Schätzung abweichen und damit keinen begünstigenden Effekt hinsichtlich einer möglichen bias-Reduktion indizieren. Dem entsprechend resultieren der Adjustierung mit dem SEM M.S.E.-Schätzungen, die mit zunehmenden Stichprobenumfang zunehmen und damit keine M.S.E.-Konsistenz indizieren.

Dem gegenüber stehen, wenn auch nicht in dem Ausmaße wie in den anderen Simulationsszenarien, auffällige bias-Reduktionen, die sich nach der Adjustierung mit dem propensity score ergeben. Hervorzuheben sind an dieser Stelle bestehende Verzerrungen, die eindeutig größer ausfallen als in den anderen Simulationsszenarien und sich erklären lassen durch die Schätzung eines linearen Effekts der Variablen ξ auf die Selektionswahrscheinlichkeit in dem verwendeten logistischen Regressionsmodell. Auch an dieser Stelle wird die zu Grunde gelegte Linearitätsannahme an den datengenerierenden Prozess verletzt - jedoch mit weniger weitreichenden Konsequenzen als in dem Strukturgleichungsmodell.

Besonders eklatante Abweichungen der spezifizierten Strukturgleichungsmodelle von den datengenerierenden Prozessen führen das dritte und das fünfte Simulationsszenario auf, deren Ergebnisse an dieser Stelle diskutiert werden soll: In dem dritten Simulationsszenario wurden in dem datengenerierenden Prozess zwei latente Variablen definiert, die durch ein lineares Modell mit einander in einer Beziehung stehen und von denen die Variable η als konfundierende Störvariable einerseits das treatment-assignment und andererseits die response-Variable Y_i bedingt. In dem modellierten SEM zur parametrischen Schätzung des Effektes τ wurde die latente Variable ξ als konfundierende Störvariable in dem Strukturmodell spezifiziert, so dass die Effekte der Variablen η bei der Schätzung des Effektes τ nicht vollständig konstant gehalten werden konnten. In Konsequenz lassen sich auffällige Verzerrungen bei der adjustierten Schätzung des Effekts τ beobachten, die auch in großen Stichproben bestehen bleiben und eine Überschätzung des Effektes aufweisen. Offenkundig ermöglicht die formulierte Korrelation zwischen beiden latenten Variablen in dem Strukturgleichungsmodell weiterhin eine verhältnismäßig geringe Konstanthaltung der Variablen η bei der Schätzung des interessierenden Effektes τ, da trotz der aufgewiesenen auffälligen Verzerrungen moderate bias-Reduktionen zu berichten sind. Andererseits ist die resultierende Verzerrung bereits in kleinen Stichproben, in denen verhältnismäßig große Standardfehler zu berichten sind, so hoch, dass in den Stichproben dieser Größe kein Schätzintervall den interessierenden Effekt überlagert. Des weiteren wurden in dieser Simulation auch Verzerrungen bei der Schätzung der Faktorladungen derjenigen Indikatorvariablen, die in dem Messmodell der latenten Variablen η zugeordnet sind, aufgewiesen und sich zurückführen lassen auf die in dem SEM modellierte Korrelation zwischen den latenten Variablen η und ξ anstelle des gerichteten Pfades, der dem datengenerierenden Prozess zu Grunde liegt.

Dem gegenüber stehen die Ergebnisse der propensity score-Stratifikation als Adjustierungsmethode, der bereits in kleinen Stichproben eine auffällige Reduktion des bias, der auf die Konfundierung durch die Variable η zurückgeführt werden kann, folgt. Trotz der resultierenden bias-Reduktionen, die einen eindeutigen Effekt hinsichtlich der Reduktion der Konfundierung bei der Schätzung des treatment-Effekts offenlegen, muss auf die bestehenden Verzerrungen hingewiesen werden, die der propensity score-Stratifikation in dieser Simulation folgen: Diese sind zwar eindeutig geringer als die der adjustierten Schätzung im SEM resultierenden, weisen dennoch eine auffällige Abweichung von dem interessierenden Effekt τ auf. Diese bestehenden Verzerrungen lassen sich unter anderem hinsichtlich des datengenerierenden Prozesses erklären

durch die definierte Korrelation zwischen den beiden latenten Variablen, die sich niederschlagen wird in einer homogenen Korrelationsmatrix, aus der eine und nicht, wie in dem datengenerierenden Prozess als zweifaktorielles Modell zu Grunde gelegt, zwei Hauptkomponenten extrahiert werden. Dem folgend wird bei der Schätzung der Selektionswahrscheinlichkeiten in dem verwendeten logistischen Regressionsmodell nur eine Hauptkomponente als Prädiktorvariable aufgenommen und es resultieren dementsprechend Informationsverluste über die beiden konfundierenden latenten Variablen, die in bestehen bleibenden Konfundierungen und entsprechenden Verzerrungen bei der Schätzung des Effektes τ resultieren.

Besonders interessante Ergebnisse ließen sich in der fünften Simulation beobachten: In dem datengenerierenden Prozess standen zwei Messfehler des faktorenanalytischen Modells in einer linearen Beziehung zu einander, zusätzlich wurde einer dieser Messfehler in das logistische Selektionsmodell aufgenommen, so dass das treatment-assignment nicht nur von der latente Variable ξ, sondern zusätzlich von einem ihrer Messfehler abhängig war. Die Ergebnisse des fehlspezifizierten Strukturgleichungsmodells, das bei der Schätzung der Effekte eine Unabhängigkeit der Messfehler untereinander sowie eine Unabhängigkeit der Messfehler von den exogenen bzw. endogenen Variablen des Modells formulierte, weisen in Konsequenz eine systematische Unterschätzung des Effekts τ auf, die sich in negativen Verzerrungen und mit zunehmenden Stichprobenumfang abnehmender Standardfehler äußert. Insbesondere die Aufnahme des Messfehlers in das logistische Selektionsmodell in dem datengenerierenden Prozess, die bei der Strukturgleichungsmodellierung nicht berücksichtigt worden ist, lässt sich für diese Unterschätzung verantwortlich machen, da bei der Schätzung des Effekts τ in dem formulierten SEM lediglich der Einfluss der latenten Variablen ξ auf den Selektionsindikator konstant gehalten werden konnte, jedoch nicht der Einfluss des Messfehlers. In Konsequenz lässt sich damit bei der Schätzung des Effekts τ weiterhin ein Teil der resultierenden Fehlervarianz bei der Modellierung des treatment-assignment auf die Varianz des Messfehlers zurückführen; die Konsequenzen dieser abweichenden Modellierung des Strukturgleichungsmodells wurden bereits aufgewiesen.

Die Ergebnisse der Schätzungen nach der Adjustierung mit dem propensity score stehen im Gegensatz zu den aufgeführten Ergebnissen der Strukturgleichungsmodellierung: Bereits in kleinen Stichproben resultieren vernachlässigbare Verzerrungen, die von dem interessierenden Effekt τ lediglich auf wenige Nachkommastellen abweichen und denen in Konsequenz eine Reduktion der Konfundierung durch die latente Störvariable ξ von ca. 99.5%

folgt. Eine Erklärungsmöglichkeit für die vorliegenden Ergebnisse stellt die spezifizierte Korrelation der beiden Messfehler in dem datengenerierenden Prozess dar, die in Folge zu einer zunehmenden beobachteten Korrelation zwischen den beiden zugehörigen Indikatorvariablen, X_1 und X_2, führt. Dem folgend führt eine zunehmende Korrelation zwischen den beiden Indikatorvariablen zu extrahierten Hauptkomponenten, die in entsprechend größeren Varianzschätzungen der latenten Variablen resultieren. Somit liegen bei der Schätzung der Selektionswahrscheinlichkeiten in dem logistischen Regressionsmodell, das zur Modellierung des treatment-assignments in Abhängigkeit von den extrahierten Hauptkomponenten genutzt wird, präzisiere Varianzschätzungen der latenten Variablen in den aufgenommenen Hauptkomponenten vor, die entsprechend bei der Schätzung der propensity scores berücksichtigt werden können. Folglich bilden die geschätzten propensity scores die zu Grunde liegende Selektionswahrscheinlichkeiten in Abhängigkeit von der latenten Variablen genauer ab und es kann gegeben der propensity score-Stratifikation, die auf den geschätzten Selektionswahrscheinlichkeiten basiert, der konfundierende Effekt der latenten Variablen innerhalb eines stratums reduziert werden.

6 Diskussion

Es wurden in der vorliegenden Arbeit die Möglichkeiten untersucht und diese durch neue Vorschläge ausgeweitet, den ACE in einer observational study durch eine Adjustierung mit dem propensity score konsistent schätzen zu können, wobei für eine Identifikation des ACE im nicht-randomisierten Fall zunächst ein *strongly ignorable treatment-assignment* als gültig angenommen werden muss. Dementsprechend setzt eine Übertragung der aufgeführten Befunde in die wissenschaftliche Praxis zunächst voraus, dass die Vorstellung eines treatment-assignments, das sich an $\vec{x}_i$ bedingt als probabilistisch auffassen lässt, plausibel ist, da der interessierende Parameter anhand der Beobachtungen nur unter der Gültigkeit eines solchen Modells, das den datengenerierenden Prozess beschreibt, identifizierbar ist. Mit der Gültigkeit der mit dem strongly ignorable treatment-assignment verbundenen Modellannahmen weisen die berichteten Ergebnisse der Simulationsstudien grundlegend auf, dass die Adjustierung mit dem propensity score einen begünstigenden Effekt bei der Schätzung des ACE mit sich bringt: Eine Vielzahl der Ergebnisse zeigt, dass sich mit der Verwendung des propensity scores eine auffällige Reduktion desjenigen bias festhalten lässt, der auf die konfundierenden Variablen zurückgeführt werden kann.

Jedoch weisen insbesondere die Ergebnisse der ersten Simulationsszenarien eine eindeutige Limitierung der propensity score-Adjustierung auf: Es ist anhand der Ergebnisse unmittelbar ersichtlich, dass eine konsistente Schätzung des ACE unter Verwendung einer der propensity score-Methoden nur möglich ist, wenn in dem Schätzmodell für den unbekannten propensity score die entscheidenden Variablen, die das treatment-assignment bedingen, berücksichtigt werden. Grundlegend zeigen die ersten Simulationsszenarien, dass der zusätzlichen Hinzunahme von Variablen in das Schätzmodell, die die Selektionswahrscheinlichkeit nicht unmittelbar bedingen, weiterhin eine bias-Reduktion folgt, die auch einer exakten Modellierung des treatment-assignments resultiert, jedoch einer Exklusion der entscheidenden Variablen bei der Modellierung - ein Szenario, das dem *omitted-variables*-Problem entspricht - auffällige Verzerrungen bei der Schätzung des ACE folgen. Demnach ist es für die wissenschaftliche Praxis unerlässlich, zu Beginn einer parametrischen Schätzung des propensity scores, beispielsweise in einem logistischen Regressionsmodell, eine konkrete Vorstellung über den Selektionsmechanimus der statistischen Einheiten zu haben oder, wie es entsprechend auch in dieser Arbeit vorgeschlagen worden ist, den propensity score non-parametrisch zu schätzen. Grundsätzlich liegen, der

parametrischen Schätzung des propensity scores folgend, deskriptive Möglichkeiten vor, mit denen in der wissenschaftlichen Praxis die Balance der Kovariablen, die der Randomisierung gefolgt wäre, geprüft werden kann (z.B., Guo & Fraser, 2010). Dieses Vorgehen ermöglicht zumindest deskriptiv, die Plausibilität des verwendeten Schätzmodells für den propensity score hinsichtlich der resultierenden Homogenität der Kovariablen zu prüfen, auch wenn an dieser Stelle keine Möglichkeit dafür vorliegt, zu überprüfen, ob ggf. entscheidende Kovariablen bei der Modellierung ausgelassen worden sind.

Eine Möglichkeit, die aufgeführten Probleme, die sich bei der Modellierung des propensity scores einstellen können, zu umgehen, besteht in einer, von der Gesamtheit der erhobenen Kovariablen ausgehenden, non-parametrischen Schätzung des propensity scores. Damit könnte das aufgewiesene omitted-variables-Problem zumindest hinsichtlich der erhobenen Variablen verhindert werden, da entsprechende Schätzmodelle in den vorliegenden Vorschlägen hierzu ohne eine Vorkenntnis ausgehend von den Beobachtungen formuliert werden; für eine spezifische Klasse stetiger Kovariablen wurde in dieser Arbeit eine non-parametrische Schätzung über den log-konkaven Dichteschätzer vorgeschlagen. Per se ergibt sich hier eine grundlegende Limitierung in der Anwendung, als dass die Kovariablen, für die diese non-parametrische Schätzung möglich ist, zum Einen Realisierungen stetiger Kovariablen, zum Anderen die zugehörigen Dichtefunktionen log-konkav sein müssen. Entsprechend ist der Anwendungsbereich, für den dieser Vorschlag gültig ist, eingeengt und es müsste bei Verletzung dieser Annahmen auf einen Alternativvorschlag, wie z.B. der *logistic boosted regression* (McCaffrey et al., 2004), zurückgegriffen werden. Auch an dieser Stelle muss darauf verwiesen werden, dass eine non-parametrische Schätzung niemals das aufgeführte omitted-variables-Problem grundlegend lösen kann: es muss in praxi stets vorausgesetzt werden, dass in den Daten sämtliche Kovariablen vorliegen, die die Selektionswahrscheinlichkeit bedingen, da andernfalls die aufgeführte Verzerrungen bei der Schätzung des ACE resultieren, die empirisch bereits bei Drake (1993) gezeigt worden sind und durch ein non-parametrisches Schätzverfahren nicht behoben werden können. Des weiteren muss in der praktischen Anwendung eines non-parametrischen Schätzverfahrens stets bedacht werden, dass dieses an große Stichproben gebunden ist, um erwartungstreue, konsistente und effiziente Punktschätzungen beobachten zu können; inwieweit sich die notwendigen Stichprobengrößen in der psychologischen Forschungspraxis realisieren lassen, kann an dieser Stelle nicht eingeschätzt werden.

Eine Einschränkung in der Generalisierbarkeit der in einer wissenschaftlichen Arbeit aufgeführten Ergebnisse liegt stets in der angewandten Methodik, aus der diese hervorkommen und von der sich auch die Methodik der Simulationsstudien, die in der vorliegenden Arbeit benutzt wurde, nicht lossagen kann: Grundlegend bieten Simulationsstudien für die gewählten Szenarien, in denen die vorgeschlagenen Schätzverfahren geprüft werden, die Möglichkeit, die Ergebnisse empirisch auf Plausibilität zu prüfen, da sich entsprechende Parameter in den Simulationsszenarien festlegen lassen, gegen die diese Schätzungen, zumindest empirisch und im Idealfall, konvergieren sollten. Mit Festlegung dieser Parameter stehen damit Werte fest, die es ermöglichen zu prüfen, ob ein bestimmtes Schätz- bzw. Adjustierungsverfahren, zumindest *tendenziell*, mit den gewählten Werten übereinstimmt. Dennoch bleiben insbesondere die Erwartungstreue und die Konsistenz, zwei Kriterien, die zumeist an eine Schätzstatistik gestellt werden, nicht in der Gänze erfassbar, da eine Simulationsstudie nur mit endlichen Stichprobenumfängen arbeiten kann und auch die Anzahl an Wiederholungen, die für ein bestimmtes Szenario festgelegt werden, endlich ist. Damit bieten die berichteten Ergebnisse in dieser Arbeit nur eine empirische Annäherung an bestimmte Sachverhalte, die sich in theoretischer Hinsicht entsprechend einstellen sollten.

Für die Generierung weiterer Forschungsfragen bietet sich insbesondere der Vorschlag der non-parametrischen Schätzung des propensity scores an: es bleibt ausgehend von der Idee, den unbekannten propensity score non-parametrisch über den log-konkaven Dichteschätzer zu schätzen, zu klären, ob sich weitere Vorschläge zur Dichteschätzung finden lassen, die von der Klasse der log-konkaven Dichten losgelöst sind und sich allgemein auf die Klasse der stetigen Kovariablen anwenden lassen. Bisher sind derartig allgemein gehaltene Vorschläge mit den aufgewiesenen Vorzügen, die der log-konkave Dichteschätzer besitzt, nicht bekannt, so dass dieser in der Klasse der Dichteschätzer zumindest gegenwärtig eine herausragende Stellung einnimmt, die sich bereits darin widerspiegelt, als dass die Veröffentlichung (Cule et al., 2010) einen Umfang von insgesamt 60 Seiten im *Journal of Royal Statistical Society: Series B* einnimmt. Mit einem allgemeinen Vorschlag zur Schätzung einer unbekannten, multivariaten Dichte, die in der Modellierung des propensity scores eine entsprechende Rolle einnimmt, ließe sich der vorliegende Vorschlag unmittelbar ausweiten auf weniger spezifische Szenarien. Bis zu dem Zeitpunkt wird die beschriebene *logistic boosted regression* zumindest in der Bandbreite möglicher Anwendungen dem vorliegenden Vorschlag eindeutig überlegen sein.

Abbildungsverzeichnis

Literatur

Abadie, A. & Imbens, G. W. (2002). *Simple and bias-corrected matching estimators for average treatment effects.* Cambridge.

Abadie, A. & Imbens, G. W. (2006). Large sample properties of matching estimators for average treatment effects. *Econometrica*, *74* (1), 235–267.

Abadie, A. & Imbens, G. W. (2011). Bias-Corrected Matching Estimators for Average Treatment Effects. *Journal of Business & Economic Statistics*, *29* (1), 1–11.

Anderson, J. C. & Gerbing, D. W. (1988). Structural equation modeling in practice: A review and recommended two-step approach. *Psychological Bulletin*, *103* (3), 411–423.

Austin, P. C. (2011). An Introduction to Propensity Score Methods for Reducing the Effects of Confounding in Observational Studies. *Multivariate behavioral research*, *46* (3), 399–424.

Baser, O. (2006). Too much ado about propensity score models? Comparing methods of propensity score matching. *Value in health: The Journal of the International Society for Pharmacoeconomics and Outcomes Research*, *9* (6), 377–385.

Bentler, P. M. (1980). Multivariate analysis with latent variables: Causal modeling. *Annual review of psychology*, *31*, 419–456.

Blundell, R. & Costa-Dias, M. (2002). *Alternative approaches to evaluation in empirical microeconomics.* London.

Bollen, K. A. (1989). *Structural equations with latent variables.* New York: John Wiley & Sons.

Bollen, K. A. (2002). Latent Variables in Psychology and the Social Sciences. *Annual review of psychology*, *53*, 605–634.

Browne, M. (1984). Asymptotically distribution-free methods for the analysis of covariance structures. *British Journal of Mathematical and Statistical Psychology*, *37* (1), 62–83.

Burns, D. D. & Nolen-Hoeksema, S. (1992). Therapeutic empathy and recovery from depression in cognitive-behavioral therapy: a structural equation model. *Journal of consulting and clinical psychology*, *60* (3), 441–449.

Campbell, D. T. (1957). Factors relevant to the validity of experiments in social settings. *Psychological Bulletin*, *54* (4), 297–312.

Chesher, A. (1991). The Effect of Measurement Error. *Biometrika*, *78* (3), 451–462.

Cochran, W. G. (1968). The effectiveness of adjustment by subclassification in removing bias in observational studies. *Biometrics*, *24* (2), 295–313.

Cochran, W. G. & Chambers, S. (1965). The planning of observational studies of human populations. *Journal of the Royal Statistical Society. Series A (General)*, *128* (2), 234–266.

Cochran, W. G. & Rubin, D. B. (1973). Controlling bias in observational studies: A review. *Sankhya: The Indian Journal of Statistics, Series A (1961-2002)*, *35* (4), 417–446.

Cox, D. (1992). Causality: some statistical aspects. *Journal of the Royal Statistical Society. Series A (Statistics in Society)*, *155* (2), 291–301.

Cule, M., Samworth, R. & Stewart, M. (2010). Maximum likelihood estimation of a multidimensional log-concave density. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, *72* (5), 545 – 607.

D' Agostino, R. B. (1998). Propensity score methods for bias reduction in the comparsion of treatment to a non-randomized control group. *Statistics in medicine*, *17*, 2265–2281.

Dawid, A. (1979). Conditional independence in statistical theory. *Journal of the Royal Statistical Society. Series B (Methodological)*, *41* (1), 1–31.

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

Dorn, H. F. (1953). Philosophy of inferences from retrospective studies. *American journal of public health and the nation's health*, *43* (6 Pt 1), 677 – 683.

Drake, C. (1993). Effects of misspecification of the propensity score on estimators of treatment effect. *Biometrics*, *49* (4), 1231–1236.

Emura, T., Wang, J. & Katsuyama, H. (2008). *Assessing the Assumption of Strongly Ignorable Treatment Assignment Under Assumed Causal Models* (Bericht).

Epanechnikov, V. (1969). Non-parametric estimation of a multivariate probability density. *Theory of Probability & Its Applications*.

Fisher, R. A. (1926). The arrangement of field experiments. *Journal of the Ministry of Agricuture of Great Britain*, *33*, 503–513.

Fisher, R. A. & Wishart, J. (1930). *The Arrangement of Field Experiments and the Statistical Reduction of the Results.* (Bd. Technical) (Nr. 10).

Fox, J. (2006). Structural Equation Modeling With the sem Package in R. *Structural Equation*

Modeling, *13* (3), 465–486.

Frölich, M. (2004). Finite-sample properties of propensity-score matching and weighting estimators. *Review of Economics and Statistics*, *86* (1), 77–90.

Guo, S. & Fraser, M. W. (2010). *Propensity Score Analysis: Statistical Methods and Applications* (Bd. 11; S. Y. Guo & M. W. Fraser, Hrsg.) (Nr. Book, Whole). Sage Publications.

Guttman, L. (1954). Some necessary conditions for common-factor analysis. *Psychometrika*, *19* (2), 149–161.

Hahn, J. (1998). On the role of the propensity score in efficient semiparametric estimation of average treatment effects. *Econometrica*, *66* (2), 315–331.

Hainmueller, J. (2011). Entropy Balancing for Causal Effects: A Multivariate Reweighting Method to Produce Balanced Samples in Observational Studies. *Political Analysis*, *20* (1), 25–46.

Hansen, B. B. (2004). Full Matching in an Observational Study of Coaching for the SAT. *Journal of the American Statistical Association*, *99* (467), 609–618.

Heckman, J. J. (1979). Sample selection bias as a specification error. *Econometrica: Journal of the econometric society*, *47* (1), 153–161.

Heckman, J. J. (1990). Varieties of selection bias. *The American Economic Review*, *80* (2), 313–318.

Heckman, J. J. (1991). *Randomization and social policy evaluation.* Cambridge.

Heckman, J. J. (2005). The scientific model of causality. *Sociological methodology*, *35* (1), 1–97.

Heckman, J. J., Ichimura, H. & Todd, P. (1997). Matching as an econometric evaluation estimator: Evidence from evaluating a job training programme. *The review of economic studies*, *64* (4), 605–654.

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

Hill, A. B. (1965). The environment and disease: association or causation? *Proceedings of the Royal Society of Medicine*, 295–300.

Hirano, K., Imbens, G. & Ridder, G. (2003). Efficient estimation of average treatment effects using the estimated propensity score. *Econometrica*, *71* (4), 1161–1189.

Höfler, M. (2005). Causal inference based on counterfactuals. *BMC Medical Research Metho-*

dology, *5* (28), 1–12.

Holland, P. W. (1986). Statistics and causal inference. *Journal of the American Statistical Association*, *81* (396), 945–960.

Holland, P. W. & Rubin, D. B. (1988). Causal Inference in Retrospective Studies. *Evaluation Review*, *12* (3), 203–231.

Horvitz, D. G. & Thompson, D. J. (1952). A generalization of sampling without replacement from a finite universe. *Journal of the American Statistical Association*, *47* (260), 663–685.

Hotelling, H. (1933). Analysis of a complex of statistical variables into principal components. *Journal of Educational Psychology*, *24* (6), 417–441.

Hume, D. (1739). *A Treatise of Human Nature*. Baltimore, MD: Penguin.

Imai, K., King, G. & Stuart, E. A. (2008). Misunderstandings between experimentalists and observationalists about causal inference. *Journal of the Royal Statistical Society. Series A (Statistics in Society)*, *171* (2), 481–502.

Imai, K. & Ratkovic, M. (2014). Covariate balancing propensity score. *Journal of the Royal Statistical Society. Series B (Statistical Methodology)*, *76* (1), 243–263.

Imbens, G. W. (2004). Nonparametric estimation of average treatment effects under exogeneity: A review. *Review of Economics and statistics*, *86* (1), 4–29.

Imbens, G. W. & Angrist, J. D. (1994). Identification and estimation of local average treatment effects. *Econometrica: Journal of the Econometric Society*, *62* (2), 467–475.

Imbens, G. W. & Rubin, D. B. (1997). Bayesian inference for causal effects in randomized experiments with noncompliance. *The Annals of Statistics*, *25* (1), 305–327.

Imbens, G. W. & Rubin, D. B. (2012). *Causal inference in statistics and social sciences*.

Johnson, R. A. & Wichern, D. W. (2007). *Applied Multivariate Statistical Analysis* (6. Aufl., Bd. 6). Upper Saddle River: Pearson Education, Inc.

Jöreskog, K. (1970). A general method for estimating a linear structural equation system. *ETS Research Bulletin Series*, *1970* (2), 1–41.

Kauermann, G. & Küchenhoff, H. (2011). *Stichproben*. Springer-Verlag Berlin Heidelberg.

Little, R. J. & Rubin, D. B. (2000). Causal effects in clinical and epidemiological studies via potential outcomes: concepts and analytical approaches. *Annual review of public health*, *21*, 121–145.

Lunceford, J. K. & Davidian, M. (2004). Stratification and weighting via the propensity score

in estimation of causal treatment effects: a comparative study. *Statistics in medicine*, *23* (19), 2937–2960.

Massy, W. F. (1965). Principal Componenets Regression in Exploratory Statistical Research. *Journal of the American Statistical Association*, *60* (309), 234–256.

McCaffrey, D. F., Ridgeway, G. & Morral, A. R. (2004). Propensity score estimation with boosted regression for evaluating causal effects in observational studies. *Psychological methods*, *9* (4), 403–25.

Morgan, S. L. (2001). Counterfactuals, causal effect heterogeneity, and the Catholic school effect on learning. *Sociology of education*, *74* (10), 341–374.

Morgan, S. L. & Winship, C. (2007). *Counterfactuals and Causal Inference: Methods and Principles for Social Research* (Bd. 88) (Nr. 1). Cambridge University Press.

Mulaik, S. A. (2009). *Linear causal modeling with structural equations* (1. Aufl.). Boca Raton: Taylor & Francis Group.

Muthén, B. & Asparouhov, T. (2011). Bayesian SEM : A more flexible representation of substantive theory. *Psychological methods*, *17* (3), 313–335.

Olsson, U. H., Foss, T., Troye, S. V. & Howell, R. D. (2000). The Performance of ML, GLS, and WLS Estimation in Structural Equation Modeling Under Conditions of Misspecification and Nonnormality. *Structural Equation Modeling*, *7* (4), 557–595.

Parzen, E. (1962). On Estimation of a Probability Density Function and Mode. *The Annals of Mathematical Statistics*, *33* (3), 1065–1076.

Pearson, K. (1901). LIII. On lines and planes of closest fit to systems of points in space. *Philosophical Magazine Series 6*, *2* (11), 559–572.

Petersen, M. L., Sinisi, S. E. & van der Laan, M. J. (2006). Estimation of direct causal effects. *Epidemiology*, *17* (3), 276–284.

Phillip, K., Gus, M., Rodney, A. & John, A. (2003). Customer repurchase intention. *European journal of marketing,*, *37* (11), 1762–1800.

Rabe-Hesketh, S., Skrondal, A. & Zheng, X. (2007). *Handbook of Latent Variable and Related Models* (1. Aufl.; S. Y. Lee, Hrsg.). Amsterdam: Elsevier.

Robins, J. M., Hernán, M. A. & Brumback, B. (2000). Marginal structural models and causal inference in epidemiology. *Epidemiology*, *11* (5), 550.

Robins, J. M., Rotnitzky, A. & Zhao, L. P. (1994). Estimation of regression coefficients when

some regressors are not always observed. *Journal of the American Statistical Association*, *89* (427), 846–866.

Rosenbaum, P. R. (1984a). The consquences of adjustment for a concomitant variable that has been affected by the treatment. *Journal of the Royal Statistical Society. Series A (General)*, *147* (5), 656–666.

Rosenbaum, P. R. (1984b). From association to causation in observational studies: The role of tests of strongly ignorable treatment assignment. *Journal of the American Statistical Association*, *79* (385), 41–48.

Rosenbaum, P. R. (1987). Model-based direct adjustment. *Journal of the American Statistical Association*, *82* (398), 387–394.

Rosenbaum, P. R. (1995). *Observational studies* (1. Aufl.). New York: Springer-Verlag.

Rosenbaum, P. R. (2010). *Design of observational studies.* New York: Springer.

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.

Rosenbaum, P. R. & Rubin, D. B. (1984). Reducing bias in observational studies using subclassification on the propensity score. *Journal of the American Statistical Association*, *79* (387), 516–524.

Rosenbaum, P. R. & Rubin, D. B. (1985a). The bias due to incomplete matching. *Biometrics*, *41* (1), 103–16.

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

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. (1976). Multivariate Matching Methods That are Equal Percent Bias Reducing, I: Some Examples. *Biometrics*, *32* (1), 109–120.

Rubin, D. B. (1977). Assignment to Treatment Group on the Basis of a Covariate. *Journal of Educational and Behavioral Statistics*, *2* (1), 1–26.

Rubin, D. B. (1978). Bayesian inference for causal effects: The role of randomization. *The Annals of Statistics*, *6* (1), 34–58.

Rubin, D. B. (2007). Statistical Inference for Causal Effects, with Emphasis on Applications in

Psychometrics and Education. In *Handbook of statistics vol 26* (Bd. 26, S. 769–800).

Rubin, D. B., Stuart, E. A. & Zanutto, E. L. (2004). A potential outcomes view of value-added assessment in education. *Journal of educational and behavioral*, *29* (1), 103–116.

Rubin, D. B. & Thomas, N. (2000). Combining Propensity Score Matching with Additional Adjustments for Prognostic Covariates. *Journal of the American Statistical Association*, *95* (450), 573–585.

Sawyer, J. E. (1992). Goal and process clarity: Specification of multiple constructs of role ambiguity and a structural equation model of their antecedents and consequences. *Journal of Applied Psychology*, *77* (2), 130–142.

Scheines, R., Hoijtink, H. & Boomsma, A. (1999). Bayesian estimation and testing of structural equation models. *Psychometrika*, *64* (1), 37–52.

Semmer, N. K., Tschan, F., Meier, L. L., Facchin, S. & Jacobshagen, N. (2010). Illegitimate Tasks and Counterproductive Work Behavior. *Applied Psychology*, *59* (1), 70–96.

Splawa-Neyman, J., Dabrowska, D. & Speed, T. (1990). On the application of probability theory to agricultural experiments. Essay on principles. Section 9. *Statistical Science*, *5* (4), 465–480.

Stuart, E. A. (2010). Matching methods for causal inference: A review and a look forward. *Statistical science : A review journal of the Institute of Mathematical Statistics*, *25* (1), 1–21.

Tan, Z. (2010). Bounded, efficient and doubly robust estimation with inverse weighting. *Biometrika*, *97* (3), 661–682.

Winship, C. & Mare, R. (1992). Models for sample selection bias. *Annual Review of Sociology*, *18*, 327–350.

Wold, H. (1956). Causal inference from observational data: A review of end and means. *Journal of the Royal Statistical Society. Series A*, *119* (1), 28–61.

Wooldridge, J. M. (2003). *Introductory Econometrics: A Modern Approach* (Bd. 2nd).

Yanai, H. & Ichikawa, M. (2007). Factor Analysis. In C. R. Rao & S. Sinharay (Hrsg.), *Handbook of statistics* (1. Aufl., Bd. 26, S. 257–296). Amsterdam: Elsevier.

Yuan, K.-J. & Bentler, P. M. (2007). Structural Equation Modelling. In C. R. Rao & S. Sinharay (Hrsg.), *Handbook of statistics* (1. Aufl., Bd. 26, S. 297–358). Amsterdam: Elsevier.

Zhao, Z. (2004). Using Matching to Estimate Treatment Effects: Data Requirements, Matching

Metrics, and Monte Carlo Evidence. *Review of Economics and Statistics*, *86* (1), 91–107.

Anhang

Programmcode der ersten Simulationsstudie

```
set.seed(500)
library(Matching)

GG <-c(1:10000)
tau <-0
beta <-10
n.ziehung <-seq(1, 5000)
n <-c(100, 200, 500)

##Kovariablen##
x1<-sample(c(0,1), length(GG), replace=T)
x2<-rnorm(length(GG), 30,5)
x3<-x2 + rnorm(length(GG), 0, 1)
x4<-sample(c(1:4), length(GG), replace=T)
x5<-x3 + rnorm(length(GG), 0, 20)
x6<-rnorm(length(GG), 2, 10))

##response##
y0 <-round(rnorm(length(GG),100,15))+beta*x1
y1.1<-y0+tau
daten<-data.frame(cbind(y0,y1.1,x1,x2,x3,x4,x5))

for(x in 1:length(n)){
for(i in 1:length(n.ziehung)){
index<-sample(c(1:length(GG)), n[x], replace=F)
newdat<-data.frame(daten[index,])

b0<-0
```

```
b1<-log(2)
t <- numeric(length(index))
     for(j in 1:length(index)){
        wkeit <- plogis(b0+b1*newdat$x1[j])
        t[j] <- ifelse(runif(1)<wkeit,1,0)}
newdat2<-data.frame(newdat, t)

##unadjustierte Schätzung##
response_kontrolle<-newdat2[t==0,1]
response_treatment<-newdat2[t==1,2]
Punktschätzer[i]<-mean(response_treatment)-mean(response_kontrolle)
Freiheitsgrade1<-(length(response_treatment)-1)
Freiheitsgrade2<-(length(response_kontrolle)-1)
Freiheitsgrade_gesamt<-(length(response_treatment)+length(response_kontrolle)-2)
var_pooled[i]<-( (Freiheitsgrade1*var(response_treatment) +
 (Freiheitsgrade2*var(response_kontrolle))) / Freiheitsgrade_gesamt)
var_pooled_mw[i]<- var_pooled[i] * ((1/length(response_treatment)) +
   (1/length(response_kontrolle)))
se_mittelwert[i]<-sqrt(var_pooled[i]*((1/length(response_treatment)) +
(1/length(response_kontrolle))))
ki_u[i]<-Punktschätzer[i]-1.96*se_mittelwert[i]
ki_o[i]<-Punktschätzer[i]+1.96*se_mittelwert[i]


##Propensity score: Schätzung##
prop.score<-NULL
Design<-cbind(newdat2$x1, newdat2$x2, newdat2$x3, newdat2$x4, newdat2$x5, newdat2$x6)
logmod<-glm(newdat2$t ~ Design, family="binomial")
prop.score<-logmod$fitted.values
```

```
##Dezentil-Stratifikation##
newdat3<-data.frame(newdat2, prop.score)
decentil<-quantile(newdat3$prop.score, probs=seq(0,100,by=10)/100)
decentil[1]<--Inf
Schätzer_prop<-NULL
SE_prop<-NULL
Gewicht<-NULL
n.n<-0

for(l in 1:(length(decentil) - 1)){
  data<-newdat3[newdat3$prop.score > decentil[l] &
  newdat3$prop.score <= decentil[l+1],]
  if(sum(data$t == 1) > 1 & sum(data$t==0) > 1){
  Schätzer_h<-(mean(data$y1.1[data$t==1]) - mean(data$y0[data$t==0]))
  Schätzer_prop<-c(Schätzer_prop, Schätzer_h)
  Gewicht<-c(Gewicht, dim(data)[1])
  n.n<-n.n+dim(data)[1]
  QS_stratum_treat<-(length(data$y1.1[data$t==1])-1) * var(data$y1.1[data$t==1])
  QS_stratum_kontroll<-(length(data$y0[data$t==0])-1) * var(data$y0[data$t==0])
  var_pooled_stratum<-((QS_stratum_treat + QS_stratum_kontroll) /
  (length(data$y1.1[data$t==1]) + length(data$y0[data$t==0]) - 2))
  varfehler_stratum<-(var_pooled_stratum*((1/length(data$y0[data$t==0]))+
  (1/length(data$y1.1[data$t==1]))))
  SE_prop<-c(SE_prop,varfehler_stratum)}}
Gewicht_5<-Gewicht/n.n
Punktschätzer_prop_strat[i]<-sum(Gewicht_5*Schätzer_prop)
var_mittelw_strat_prop[i]<-sum(SE_prop*Gewicht_5^2)
se_prop_strat[i]<-sqrt(var_mittelw_strat_prop[i])
ki_u_strat_prop[i]<-Punktschätzer_prop_strat[i]-1.96*se_prop_strat[i]
ki_o_strat_prop[i]<-Punktschätzer_prop_strat[i]+1.96*se_prop_strat[i]
```

```
##Quintil-Stratifikation##
quintile<-quantile(newdat3$prop.score, probs=seq(0, 1, by=0.2))
quintile[1]<--Inf
Schätzer_prop_quin<-NULL
SE_prop_quin<-NULL
Gewicht_quin<-NULL
n.n_quin<-0

for(l in 1:(length(quintile) - 1)){
  data<-newdat3[newdat3$prop.score > quintile[l] &
  newdat3$prop.score <= quintile[l+1],]
  if(sum(data$t==1) > 1 & sum(data$t == 0) > 1){
  Schätzer_h<-(mean(data$y1.1[data$t==1]) - mean(data$y0[data$t==0]))
  Schätzer_prop_quin<-c(Schätzer_prop, Schätzer_h)
  Gewicht_quin<-c(Gewicht_quin, dim(data)[1])
  n.n_quin<-n.n_quin + dim(data)[1]
  QS_stratum_treat<-(length(data$y1.1[data$t==1])-1) * var(data$y1.1[data$t==1])
  QS_stratum_kontroll<-(length(data$y0[data$t==0])-1) * var(data$y0[data$t==0])
  var_pooled_stratum<-((QS_stratum_treat + QS_stratum_kontroll) /
   (length(data$y1.1[data$t==1]) + length(data$y0[data$t==0]) - 2))
  varfehler_stratum<-(var_pooled_stratum*((1/length(data$y0[data$t==0]))
   +(1/length(data$y1.1[data$t==1]))))
  SE_prop_quin<-c(SE_prop,varfehler_stratum)}}
Gewicht_6<-Gewicht_quin/n.n_quin
Punktschätzer_quintile[i]<-
sum(Schätzer_prop_quin*rep(Gewicht_6[1],length(Schätzer_prop_quin)))
Var_quintile[i]<-sum(SE_prop_quin*rep(Gewicht_6[1]^2, length(SE_prop_quin)))
Sd_quintile[i]<-sqrt(Var_quintile[i])
ki_quin_u[i]<-Punktschätzer_quintile[i] - 1.96*Sd_quintile[i]
ki_quin_o[i]<-Punktschätzer_quintile[i] + 1.96*Sd_quintile[i]
```

```
##propensity weighting 1##
Z<-newdat3$t
ZY<- (Z*newdat3$y1.1) + ((1-Z)*newdat3$y0)
e<-newdat3$prop.score
tau_weight1[i]<-((1/n[x])*sum(Z*ZY / e) - (1/n[x])*sum((1-Z)*ZY/(1-e)))
Delta_1<-tau_weight1[i]
W<-cbind(1, Design)
Summe<-0
Summe2<-0
for(k in 1:n[x]){
Summe <- Summe + (((Z[k]*ZY[k]*(1-e[k]))/e[k]) +
(((1-Z[k]) * ZY[k] *e[k])/ (1-e[k])))*W[k,]
Summe2 <- Summe2 + (e[k]* (1-e[k])) * matrix(W[k,], ncol=1)
%*% matrix(W[k,], nrow=1)}
H_beta_1<- (1/n[x]) * Summe
E_beta_1<-(1/n[x])*Summe2
Summe3<-c(rep(0,n[x]))
for(k in 1:n[x]){
     Summe3[k] <- (Z[k] * ZY[k])/e[k] - ((1-Z[k]) * ZY[k])/(1-e[k]) - Delta_1 -
     (Z[k] - e[k])* matrix(H_beta_1, nrow=1) %*% solve(E_beta_1)
     %*% matrix(W[k,], ncol=1)}
Varianz_weight_1[i]<- sum(Summe3^2)/n[x]^2
se_weight_1[i]<-sqrt(Varianz_weight_1[i])
ki_u_weight_1[i]<- tau_weight1[i]-1.96*se_weight_1[i]
ki_o_weight_1[i]<- tau_weight1[i]+1.96*se_weight_1[i]

##propensity weighting 2##
mu_weight_1<-(1/sum(Z/e))* sum(Z*ZY/e)
mu_weight_0<- (1/(sum((1-Z)/(1-e))) * sum(((1-Z)*ZY)/(1-e)))
tau_weight_2[i]<- mu_weight_1 - mu_weight_0
summe2<-0
```

```
for(k in 1:n[x]){
summe2<- summe2 + (((Z[k]*(ZY[k] - mu_weight_1)*(1-e[k]))/e[k])
+ (((1-Z[k])*(ZY[k] - mu_weight_0)*e[k])/(1-e[k])))*W[k,]}
H_beta_2<-(1/n[x])*summe2
E_beta_2<-E_beta_1
summe_22<-c(rep(0,n[x]))
for(k in 1:n[x]){
summe_22[k]<- ((Z[k]*(ZY[k] - mu_weight_1)) / e[k]) - (((1-Z[k])*(ZY[k]-
mu_weight_0))/(1-e[k])) -
(Z[k] - e[k])* matrix(H_beta_2, nrow=1) %*%
solve(E_beta_2) %*% matrix(W[k,], ncol=1)}
Varianz_weight_2[i]<- sum(summe_22^2)/n[x]^2
SE_weight_2[i]<-sqrt(Varianz_weight_2[i])
ki_u_weight2[i]<-tau_weight_2[i] - 1.96*SE_weight_2[i]
ki_o_weight2[i]<-tau_weight_2[i] + 1.96*SE_weight_2[i]

##matching##
X<-newdat3$prop.score
Y<- ZY
Tr<-Z
rr<-Match(Y=Y, Tr=Tr, X=X, M=1)
tau_matched[i]<-rr$est
sd_matched[i]<-rr$se
var_matched[i]<-(rr$se)^2
ki_match_u[i]<- tau_matched[i]-1.96*sd_matched[i]
ki_match_o[i]<- tau_matched[i]+1.96*sd_matched[i]}}
```

Programmcode der zweiten Simulationsstudie

```
set.seed(100)
library(Matching)
library(mvtnorm)
library(LogConcDEAD)

GG<-c(1:10000)
tau<-0
beta<-10

X<-rnorm(length(GG))
y0<-rnorm(length(GG),100,15)+beta*X
y1.1<-y0+tau
daten<-data.frame(cbind(y0,y1.1,X))

n.wd<-400
n<-250
for(i in 1:n.wd){
index<-sample(length(GG), n , replace=F)
SP<-daten[index,]

###treatment-Selektion
b0<-0
b1<-log(1.5)
t <- numeric(length(index))
for(j in 1:length(index)){
       wkeit <- plogis(b0 + b1*SP$X[j])
       t[j] <- ifelse(runif(1) < wkeit,1,0)}
newdat<-data.frame(SP, t)

##Punktschätzung: unadjustiert##
```

```
response_kontrolle<-newdat[newdat$t == 0,1]
response_treatment<-newdat[newdat$t == 1,2]
Punkt_1[i]<-(mean(response_treatment)-mean(response_kontrolle))
QS_t<-(length(response_treatment) - 1) * var(response_treatment)
QS_k<-(length(response_kontrolle) - 1) * var(response_kontrolle)
Var_unad_1[i]<- ((QS_t + QS_k) / (length(response_treatment) +
 length(response_kontrolle) -2))
SE_unad_1[i] <-sqrt(((1/length(response_kontrolle)) +
(1/length(response_treatment))) * Var_unad_1[i])
KI_U_unad_1[i]<- Punkt_1[i] - 1.96*SE_unad_1[i]
KI_O_unad_1[i]<- Punkt_1[i] + 1.96*SE_unad_1[i]

##Adjustierungen##
#wahrer propensity score
treat<-newdat$t
prop.score<-NULL
Design<-cbind(newdat$X)
logmod<-glm(treat ~ Design, family="binomial")
prop.score<-logmod$fitted.values
newdat<-data.frame(newdat, prop.score)

#log-konkaver Dichteschätzer
treatment<-newdat[newdat$t==1,]
Kontrolle<-newdat[newdat$t==0,]

#Für treatment#
XX<-cbind(treatment$X)
MLE<-exp(mlelcd(XX)$logMLE)

#Für Kontrolle#
XY<-cbind(Kontrolle$X)
```

```
MLE2<-exp(mlelcd(XY)$logMLE)

##Bayes##
pr_x1_s1<-MLE
pr_x1_s0<-dlcd(XX, mlelcd(XY),uselog=F)
pr_x0_s0<-MLE2
pr_x0_s1<-dlcd(XY, mlelcd(XX),uselog=F)
pr_treat<-as.vector(prop.table(table(newdat$t)))[2]
pr_kont<-as.vector(prop.table(table(newdat$t)))[1]

#treatment-Gruppe:Dichteschätzung#
Bayes_treat<-(pr_x1_s1 * pr_treat)/((pr_x1_s1 * pr_treat)+(pr_x1_s0 * pr_kont))
#Kontrollgruppe:Dichteschätzung#
Bayes_kont<-(pr_x0_s1 * pr_treat)/((pr_x0_s1 * pr_treat)+(pr_x0_s0 * pr_kont))

index2<-newdat$t == 1
propscore<-rep(0, n)
propscore[index2]<-Bayes_treat
propscore[!index2]<-Bayes_kont
newdat<-data.frame(newdat, propscore)

##ACE-Schätzung mit echten prop.-score##
quintile<-quantile(newdat$prop.score, probs=seq(0, 1, by=0.2))
quintile[1]<--Inf
Schätzer_prop_quin<-NULL
Var_prop_quin<-NULL
Gewicht_quin<-NULL
SE_prop_quin<-NULL
n.n_quin<-0

for(l in 1:(length(quintile) - 1)){
```

```
  data<-newdat[newdat$prop.score > quintile[l] & newdat$prop.score <= quintile[l+1],]
      if(sum(data$t==1) > 1 & sum(data$t == 0) > 1){
        Schätzer_h<-(mean(data$y1.1[data$t==1]) - mean(data$y0[data$t==0]))
        Schätzer_prop_quin<-c(Schätzer_prop_quin, Schätzer_h)
        Gewicht_quin<-c(Gewicht_quin, dim(data)[1])
        n.n_quin<-n.n_quin + dim(data)[1]
        ##Varianzschätzung##
        QS_stratum_treat<-(length(data$y1.1[data$t==1])-1) * var(data$y1.1[data$t==1])
        QS_stratum_kontroll<-(length(data$y0[data$t==0])-1) * var(data$y0[data$t==0])
        var_pooled_stratum<-((QS_stratum_treat + QS_stratum_kontroll) /
         (length(data$y1.1[data$t==1]) + length(data$y0[data$t==0]) - 2))
        varfehler_stratum<-(var_pooled_stratum*((1/length(data$y0[data$t==0]))
         + (1/length(data$y1.1[data$t==1]))))
        Var_prop_quin<-c(SE_prop_quin,varfehler_stratum)}}
Gewicht_6<-Gewicht_quin/n.n_quin
Punktschätzer_quintile_1[i]<-sum(Schätzer_prop_quin*rep(Gewicht_6[1],
length(Schätzer_prop_quin)))
Var_MW_quintile_1[i]<-sum((Gewicht_6^2)*Var_prop_quin)
SE_MW_quintile_1[i]<-sqrt(Var_MW_quintile_1[i])
KI_echter_U_1[i]<-Punktschätzer_quintile_1[i] - 1.96*SE_MW_quintile_1[i]
KI_echter_O_1[i]<-Punktschätzer_quintile_1[i] + 1.96*SE_MW_quintile_1[i]

##ACE-Schätzung mit logkonkav geschätzten prop.-score##
quintile_log<-quantile(newdat$propscore, probs=seq(0, 1, by=0.2))
quintile_log[1]<--Inf
Schätzer_prop_log_quin<-NULL
Var_prop_log_quin<-NULL
Gewicht_quin_log<-NULL
n.n_quin_log<-0
for(l in 1:(length(quintile_log) - 1)){
```

```
  data_3<-newdat[newdat$propscore > quintile_log[l] &
   newdat$propscore <= quintile_log[l+1],]
      if(sum(data_3$t==1) > 1 & sum(data_3$t == 0) > 1){
           Schätzer_h_log<-(mean(data_3$y1.1[data_3$t==1]) -
            mean(data_3$y0[data_3$t==0]))
           Schätzer_prop_log_quin<-c(Schätzer_prop_log_quin, Schätzer_h_log)
           Gewicht_quin_log<-c(Gewicht_quin_log, dim(data_3)[1])
           n.n_quin_log<-n.n_quin_log + dim(data_3)[1]

           QS_stratum_treat<-(length(data_3$y1.1[data_3$t==1])-1) *
           var(data_3$y1.1[data_3$t==1])
           QS_stratum_kontroll<-(length(data_3$y0[data_3$t==0])-1) *
           var(data_3$y0[data_3$t==0])
           var_pooled_stratum<-((QS_stratum_treat + QS_stratum_kontroll) /
           (length(data_3$y1.1[data_3$t==1]) + length(data_3$y0[data_3$t==0]) - 2))
           varfehler_stratum<-(var_pooled_stratum*((1/length(data_3$y0[data_3$t==0]))
            +(1/length(data_3$y1.1[data_3$t==1]))))
           Var_prop_log_quin<-c(Var_prop_log_quin, varfehler_stratum)}}
Gewicht_6_log<-Gewicht_quin_log/n.n_quin_log
Punktschätzer_log_quintile_1[i]<-sum(Schätzer_prop_log_quin*rep(Gewicht_6_log[1],
 length(Schätzer_prop_log_quin)))
Var_MW_quintile_log_1[i]<-sum((Gewicht_6_log^2)*Var_prop_log_quin)
SE_MW_quintile_log_1[i]<-sqrt(Var_MW_quintile_log_1[i])
KI_log_U_1[i]<-Punktschätzer_log_quintile_1[i] - 1.96*SE_MW_quintile_log_1[i]
KI_log_O_1[i]<-Punktschätzer_log_quintile_1[i] + 1.96*SE_MW_quintile_log_1[i]

##Differenzen##
log_diff_1[i]<-mean(abs(newdat$prop.score-newdat$propscore))}
```

Programmcode der dritten Simulationsstudie

```
set.seed(100)
library(polycor)
library(sem)

n.wd  <- 500
n     <- c(200,500,1000)

for(m in 1:length(n)){
for(i in 1:n.wd){
#Faktorenanalytisches Modell
F1 <- rnorm(n[m], 0, 1)

#Indikatorvariablen mit Messfehler
X1 <- 1.8*F1 + rnorm(n[m], 0, 1)
X2 <- 1.5*F1 + rnorm(n[m], 0, 1)
X3 <- 1.6*F1 + rnorm(n[m], 0, 1)
X4 <- 1.5*F1 + rnorm(n[m], 0, 1)
X5 <- 1.2*F1 + rnorm(n[m], 0, 1)
X6 <- 1.5*F1 + rnorm(n[m], 0, 1)
X7 <- 1.4*F1 + rnorm(n[m], 0, 1)
X8 <- 1.3*F1 + rnorm(n[m], 0, 1)
X  <-  cbind(X1, X2, X3, X4, X5, X6, X7, X8)

#response-Variablen
tau  <- 2
beta <- 10
y0 <- rnorm(n[m], 0, 1) + beta*F1
y1 <- y0 + tau
newdat <- data.frame(y0, y1, F1, X)
```

```
#Selektionsmodell
    b0 <- 0
    b1 <- log(3)
    t  <- NULL
        for(l in 1:n[m]){
          wkeit <- plogis(b0 + b1*newdat$F1[l])
          t[l] <- ifelse(runif(1) < wkeit, 1, 0)}
    t      <- factor(t, levels=c(0,1))
    newdat <- data.frame(newdat, t)

    Y    <- NULL
      for(j in 1:n[m]){Y[j] <- ifelse(newdat$t[j] == 0, newdat$y0[j], newdat$y1[j])}
      newdat <- data.frame(newdat, Y)
##ACE - Schätzung##
##unadjustierte Schätzung##
response_treatment    <- newdat$Y[newdat$t == 1]
response_kontrolle    <- newdat$Y[newdat$t == 0]
Punktschätzer[i]      <- mean(response_treatment) - mean(response_kontrolle)
Freiheitsgrade1       <-(length(response_treatment)-1)
Freiheitsgrade2       <-(length(response_kontrolle)-1)
Freiheitsgrade_gesamt <-(length(response_treatment)+length(response_kontrolle)-2)
var_pooled[i]   <-( (Freiheitsgrade1*var(response_treatment)  +
 (Freiheitsgrade2*var(response_kontrolle))) / Freiheitsgrade_gesamt)
var_pooled_mw[i]<- var_pooled[i]*((1/length(response_treatment))
  + (1/length(response_kontrolle)))
se_mittelwert[i]<- sqrt(var_pooled[i]*((1/length(response_treatment))
  + (1/length(response_kontrolle))))
ki_u[i]         <- Punktschätzer[i]-1.96*se_mittelwert[i]
ki_o[i]         <- Punktschätzer[i]+1.96*se_mittelwert[i]

##SEM
```

```
Dat <- data.frame(newdat$Y, newdat$t, newdat$X1, newdat$X2, newdat$X3, newdat$X4,
newdat$X5, newdat$X6, newdat$X7, newdat$X8)
colnames(Dat) <- c("Y", "t", "X1", "X2", "X3", "X4", "X5", "X6", "X7", "X8")

#unstandardisiert
Sigma <- var(Dat)
modell1 <- specifyModel(file="C:/Users/Marco G/Desktop/Simulation 3/
Modellspezifikation.r")
sem.test <- sem(modell1, Sigma, N = n[m])

#treatment-Effekt
PunktschätzerSEM[i]  <- sem.test$coeff[names(sem.test$coeff) == "tau"]
StandardfehlerSEM[i] <- summary(sem.test)$coeff[
rownames(summary(sem.test)$coeff) == "tau",2]
ki_u_sem[i]          <- PunktschätzerSEM[i]-1.96*StandardfehlerSEM[i]
ki_o_sem[i]          <- PunktschätzerSEM[i]+1.96*StandardfehlerSEM[i]
chi                  <- summary(sem.test)$chisq
df                   <- summary(sem.test)$df
Sign[i]              <- 1-pchisq(chi, df=df)

#Propensity score
#Eigenwertszerlegung
Kovariablen <- data.frame(newdat$X1,newdat$X2,newdat$X3,newdat$X4,
newdat$X5, newdat$X6, newdat$X7, newdat$X8)
PCA <- princomp(Kovariablen, cor=T)
Hauptkomponenten <- PCA$scores[,PCA$sdev >= 1]
newdat <- data.frame(newdat, Hauptkomponenten)

prop.score<-NULL
Design<-Hauptkomponenten
logmod<-glm(newdat$t ~ Design, family="binomial")
```

```
prop.score<-logmod$fitted.values

#Dezentilstratifikation
newdat3 <- data.frame(newdat, prop.score)
decentil<-quantile(newdat3$prop.score, probs=seq(0,100,by=10)/100)
decentil[1]<--Inf
Schätzer_prop<-NULL
SE_prop<-NULL
Gewicht<-NULL
n.n<-0
for(k in 1:(length(decentil) - 1)){
data <- newdat3[newdat3$prop.score > decentil[k] &
newdat3$prop.score <= decentil[k+1],]
if(sum(data$t == 1) > 1 & sum(data$t==0) > 1){
Schätzer_h<-(mean(data$Y[data$t==1]) - mean(data$Y[data$t==0]))
Schätzer_prop<-c(Schätzer_prop, Schätzer_h)
Gewicht<-c(Gewicht, dim(data)[1])
n.n<-n.n+dim(data)[1]
QS_stratum_treat<-(length(data$Y[data$t==1])-1) * var(data$Y[data$t==1])
QS_stratum_kontroll<-(length(data$Y[data$t==0])-1) * var(data$Y[data$t==0])
var_pooled_stratum<-((QS_stratum_treat + QS_stratum_kontroll) /
(length(data$Y[data$t==1]) + length(data$Y[data$t==0]) - 2))
varfehler_stratum<-(var_pooled_stratum*((1/length(data$Y[data$t==0]))+
(1/length(data$Y[data$t==1]))))
SE_prop<-c(SE_prop,varfehler_stratum)}}
Gewicht_5<-Gewicht/n.n
Punktschätzer_Dezentile[i]<-sum(Gewicht_5*Schätzer_prop)
var_mw_Dezentile[i]       <-sum(SE_prop*Gewicht_5^2)
se_Dezentile[i]           <-sqrt(var_mw_Dezentile[i])
ki_u_Dezentile[i]         <-Punktschätzer_Dezentile[i]  - 1.96 * se_Dezentile[i]
ki_o_Dezentile[i]         <-Punktschätzer_Dezentile[i]  + 1.96 * se_Dezentile[i]}
```

Universität Hamburg
DER FORSCHUNG | DER LEHRE | DER BILDUNG

Fakultät für
Psychologie und
Bewegungswissenschaft

Institut für Bewegungswissenschaft
Institut für Psychologie

Erklärung gemäß *(bitte Zutreffendes ankreuzen)*

☐ **§ 4 (1c) der Promotionsordnung**

des Instituts für Bewegungswissenschaft der Universität Hamburg vom 18.08.2010

☐ **§ 5 (4d) der Promotionsordnung**

des Instituts für Psychologie der Universität Hamburg vom 20.08.2003

Hiermit erkläre ich,

_____ (Vorname, Nachname),

dass ich mich an einer anderen Universität oder Fakultät noch keiner Doktorprüfung unterzogen oder mich um Zulassung zu einer Doktorprüfung bemüht habe.

_____ _____

Ort, Datum Unterschrift

Universität Hamburg
DER FORSCHUNG | DER LEHRE | DER BILDUNG

Fakultät für
Psychologie und
Bewegungswissenschaft

Institut für Bewegungswissenschaft
Institut für Psychologie

Eidesstattliche Erklärung nach *(bitte Zutreffendes ankreuzen)*

☐ **§ 7 (4) der Promotionsordnung des Instituts für Bewegungswissenschaft der Universität Hamburg vom 18.08.2010**

☐ **§ 9 (1c und 1d) der Promotionsordnung des Instituts für Psychologie der Universität Hamburg vom 20.08.2003**

Hiermit erkläre ich an Eides statt,

1. dass die von mir vorgelegte Dissertation nicht Gegenstand eines anderen Prüfungsverfahrens gewesen oder in einem solchen Verfahren als ungenügend beurteilt worden ist.
2. dass ich die von mir vorgelegte Dissertation selbst verfasst, keine anderen als die angegebenen Quellen und Hilfsmittel benutzt und keine kommerzielle Promotionsberatung in Anspruch genommen habe. Die wörtlich oder inhaltlich übernommenen Stellen habe ich als solche kenntlich gemacht.

_____ _____

Ort, Datum Unterschrift