Eine Simulationsstudie zu den Grenzen der IPW Methode

Sascha Meyen, 6204566

21. September 2013

Bachelorarbeit im Studiengang Mensch-Computer-Interaktion
Universität Hamburg
Prof. Dr. Martin Spieß (Fachbereich Psychologie)
Dr. Carola Eschenbach (Fachbereich Informatik)

Zusammenfassung

Gewichtungsmethoden und speziell Inverse Probability Weighting (IPW) kann angewendet werden, wenn es in Stichproben zu systematischen Nichtbeobachtungen kommt. Der Einsatz von IPW ist jedoch nicht immer günstig, weil die Effizienz der Schätzer reduziert wird. Um heuristische Hinweise zu erarbeiten, wann IPW eingesetzt werden sollte, wird das Verhalten der Methode in einer Simulationsstudie untersucht, bei der entscheidenden Modellannahmen systematisch verletzt werden. Die Ergebnisse deuten darauf hin, dass die korrigierten Determinationskoeffizienten der geschätzten Selektions- und Regressionsmodelle mindestens kleine Zusammenhänge ($R^2 > 0.3$) zeigen sollten. Der Mittelwert der Gewichte sollte kleiner 1 und der Stichprobenumfang größer 100 sein. Unter diesen Umständen wird der Einsatz von IPW angesichts einer Nichtbeobachtungsrate von bis zu 30% empfohlen, um trotz Effizienzverlust eine angemessen Verzerrungsreduktion zu bewirken.

Inhaltsverzeichnis

1 Modellformulierung

In empirischen Untersuchungen kann der Zusammenhang einer unabhängigen Variable X und einer abhängigen Variable Y untersucht werden. Dazu werden $i = 1..N$ Datensätze $X_i, Y_i \in W_i$ in einer Stichprobe mit N Untersuchungseinheiten erhoben. Der Zusammenhang zwischen X und Y und andere Populationsprobleme, wie z.B. der Mittelwert der Variable Y, können im Allgemeinen als Minimierungsproblem für eine gegebene Stichprobe formuliert werden (Wooldridge, 2002).

$$\min_{\theta \in \Theta} \sum_{i=1}^{N} q(w_i, \theta) \tag{1}$$

Hier existiert in den meisten Anwendungsfällen ein Parameter θ_0, welcher das Minimierungsproblem löst. Zum Beispiel kann θ in der Form $q(w_i, \theta) = (y_i - x_i\theta)^2$ als der Effekt von X auf Y interpretiert werden und würde mit der Methode der kleinsten Quadrate geschätzt.

Damit die Schätzung der interessierenden Parameter unverzerrt ist, muss die Stichprobe repräsentativ sein, also w_i muss unabhängig und identisch verteilt sein wie W in der Grundgesamtheit. Die Ziehung einer repräsentativen Stichprobe bezüglich einer zu definierenden Grundgesamtheit ist ein komplexes Problem (Bortz & Döring, 2006).

Ist eine repräsentative Stichprobe gewonnen, kann es immer noch vorkommen, dass einzelne Variablen unbeobachtet bleiben (item nonresponse). Zum Beispiel wird y_i nicht beobachtet, wenn in einer Umfrage die i-te Versuchsperson keine Antwort auf Frage nach dem Einkommen Y gibt. Im Folgenden wird die Situation betrachtet, wenn es bereits gelungen ist, eine repräsentative Stichprobe zu ziehen, aber gegebenenfalls die abhängige Variable Y unbeobachtet bleibt.

Wenn Variablen der Untersuchungseinheit i fehlen, so kann im Allgemeinen $q(w_i, \theta)$ nicht berechnet werden. Ein möglicher Ansatz ist, die gezogene Stichprobe auf diejenigen Untersuchungseinheiten zu reduzieren, für die alle Variablen beobachtet sind. Dieser Ansatz wird mit Complete Case Analysis (CCA) bezeichnet (Seaman, White, Copas & Li, 2012). CCA betrachtet für jede Untersuchungseinheit i die Variable $s_i = 1$, wenn w_i vollständig beobachtet ist, sonst $s_i = 0$. Nun kann (1) umgeschrieben werden, indem alle nicht vollständig beobachteten Untersuchungseinheiten ignoriert werden.

$$\min_{\theta \in \Theta} \sum_{i=1}^{N} s_i q(w_i, \theta_{cca}) \tag{2}$$

Dieser Ansatz ergibt nur dann unverzerrte Ergebnisse, wenn das Fehlen von Variablen unabhängig von W ist, also formal $P(S = 1|W) = P(S = 1)$ (missing completely at random, MCAR). Diese Annahme gilt nicht, wenn z.B. das Einkommen Y mit höherer Wahrscheinlichkeit verschwiegen wird, wenn es sehr hoch ist. Eine Einschränkung der MCAR Annahme stellt die Annahme dar, dass die Beobachtungswahrscheinlichkeit unabhängig von Y bei gegebenem, immer beobachteten X ist (missing at random, MAR), formal $P(S = 1|W) = P(S = 1|X, Y) = P(S = 1|X)$. In diesem Fall ließe sich die Beobachtungswahrscheinlichkeit jeder Untersuchungseinheit $\pi_i = P(s_i = 1|X = x_i)$ schätzen. Dies ist nicht mehr möglich, wenn die Beobachtungswahrscheinlichkeit zusätzlich auf Y bedingt ist, also $\pi_i = P(S = 1|X, Y) \neq P(S = 1|X)$, weil in diesem Fall die Variable Y zur Vorhersage von π_i fehlt (not missing at random, NMAR). Für eine genauere Betrachtung dieser Konzeption bietet sich die Lektüre von Schafer und Graham (2002) an.

Im Allgemeinen genügt es also nicht, selbst wenn eine repräsentative Stichprobe gewonnen ist, ein systematisches Fehlen von Variablen zu ignorieren. Statt CCA bieten sich die Verfahrensklassen Multiple Imputation (MI), Gewichtung von vollständig beobachteten Untersuchungseinheiten und modellbasierte Verfahren an.

Grob zusammengefasst ersetzt MI fehlende Werte mit einer oder mehreren Schätzungen. Gewichtungsmethoden werten den Einfluss jeder vollständig beobachteten Untersuchungseinheit invers zur Beobachtungswahrscheinlichkeit, sodass seltene Datensätze höher gewichtet werden. Modellbasierte Verfahren verwenden z.B. die maximum likelihood Methode und werden hier nicht weiter besprochen. Bei Imputationsverfahren müssen Verteilungsannahmen getroffen werden, um fehlende Variablen zu schätzen. Diese Annahmen bieten zusätzliche Informationen, wodurch die Effizienz der Schätzung erhöht wird. Es kann jedoch nicht im Allgemeinen überprüft werden, ob die getroffenen Annahmen korrekt sind. MI ist dann angemessen, wenn einzelne Variablen verstreut im Datensatz fehlen (Freedman & Berk, 2008). Wenn für mehrere Untersuchungseinheiten die gleichen Teilabschnitte der Beobachtung fehlen und die immer beobachteten Variablen zur Vorhersage der Nichtbeobachtung genügen, so bieten sich Gewichtungsmethoden an. Das ist z.B. der Fall, wenn Y_i ein hochdimensionaler Vektor ist und mit MI viele fehlende Werte geschätzt würden und der immer beobachtete Vektor X_i zur Schätzung der Beobachtungswahrscheinlichkeit π_i genügt. Dieser Ansatz wird als vergleichsweise intuitiv betrachtet (Seaman et al., 2012). Beispielsweise sei eine repräsentative Stichprobe gegeben und dem Datenanalysten bekannt ist, dass nur die Hälfte der Versuchspersonen ihr Einkommen angeben, wenn es sehr hoch ist.

Nun könnten die Versuchspersonen mit hohem Einkommen doppelt gewichten werden. Dieser Ansatz findet auch Anwendung bei zensierten Daten, wie sie oft bei medizinischen Untersuchungen gefunden werden (Curtis, Hammill, Eisenstein, Kramer & Anstrom, 2007). Es ist ebenso möglich durch Gewichtung kausale Effekte in nicht experimentellen Studien zu schätzen (McClintock, 2010), wobei dafür besondere Vorsicht geboten ist.

Die vorliegende Arbeit beschäftigt sich mit einer speziellen Variante von Gewichtung, der Inverse Probability Weighting Methode (IPW). IPW verwendet die Wahrscheinlichkeit π_i, dass eine Untersuchungseinheit i vollständig beobachtet wird, und gewichtet jede Untersuchungseinheit durch das Inverse dieser Wahrscheinlichkeit π_i^{-1}. In der Literatur wird empfohlen, diese Gewichte zu stabilisieren, indem π_i mit der mittleren Beobachtungswahrscheinlichkeit $\pi = N^{-1} \sum \pi_i$ multipliziert wird, sodass die Gewichte die Form $\pi\pi_i^{-1}$ haben (Wal & Geskus, 2011). Ein gut beschriebenes, reales Anwendungsbeispiel für IPW bietet Haapea et al. (2011). Für die weiteren Betrachtungen legen wir dementsprechend folgendes Modell für IPW zugrunde.

$$\min_{\theta \in \Theta} \sum_{i=1}^{N} s_i \frac{\pi}{\pi_i} q(w_i, \theta_{ipw}) \tag{3}$$

Diese Modellformulierung liefert unverzerrte Ergebnisse unter den folgenden Annahmen (Wooldridge, 2002).

(i) Die erhobenen Untersuchungseinheiten sind repräsentativ für die Grundgesamtheit, also w_i unabhängig und identisch verteilt wie W.

(ii) w_i ist vollständig beobachtet gdw. $s_i = 1$. Es gibt eine immer beobachtete Teilmenge $v_i \in w_i$, sodass die Wahrscheinlichkeit für eine vollständige Beobachtung der i-ten Untersuchungseinheit $\pi_i = P(s_i = 1|w_i) = P(s_i = 1|v_i)$ bekannt ist oder ausreichend genau geschätzt werden kann. Die Wahrscheinlichkeiten für eine vollständige Beobachtung $\forall i : P(s_i = 1|w_i) \geq \delta > 0$ müssen für alle Untersuchungseinheiten positiv und von 0 weg gebunden sein.

Es sind zusätzlich weitere Annahmen erforderlich, die gewöhnlich gelten und hier nicht weiter untersucht werden. Für eine detailliertere Beschreibung sei auf Wooldridge (2002) verwiesen. Sind diese Annahmen erfüllt, so konvergiert für $N \to \infty$ auch die gewichtete Schätzung aus (3) gegen die unverzerrte Lösung des Populationsproblems $\theta_w \to \theta_0$.

Im Folgenden wird die Fragestellung untersucht, wann es aus praktischer Sicht empfehlenswert ist, IPW einzusetzen.

2 Fragestellung

Gewichtungsmethoden allgemein, und damit IPW im Speziellen, versuchen Verzerrung auf Kosten von Effizienz zu reduzieren (Wooldridge, 2007). Die Effizienz der Schätzer sinkt in Abhängigkeit zu den verwendeten Gewichten. Mit IPW werden Gewichte invers zu den Beobachtungswahrscheinlichkeiten $P(s_i = 1|w_i) = \pi_i$ konstruiert. Sind die Beobachtungswahrscheinlichkeiten sehr klein, so sind die konstruierten Gewichte sehr groß, wodurch die Effizienz der Schätzer sinkt. Der Grund hierfür ist, dass einzelne Untersuchungseinheiten, und damit deren Fehler, großen Einfluss auf die Schätzung erhalten. Nach Annahme (ii) sind die Beobachtungswahrscheinlichkeiten nach unten beschränkt mit $\forall i : \pi_i \geq \delta > 0$. Für $N \to \infty$ genügt die Existenz einer solchen Schranke. Für praktisches Zwecke und ein begrenztes N spielt die konkrete Ausprägung von δ jedoch eine wichtige Rolle. Da π_i in den meisten Fällen geschätzt werden muss, sind immer beobachtete Variablen für diese Schätzung notwendig. Der Einsatz von IPW ist daher zu empfehlen, wenn die Beobachtungswahrscheinlichkeiten „annehmbar“ begrenzt sind und mit den gegebenen Variablen „ausreichend“ gut geschätzt werden können. In dieser Ausarbeitung wird die Fragestellung untersucht, wann der Einsatz von IPW anhand dieser informell beschrieben Kriterien zu empfehlen ist.

Es erscheint schwierig, dieser Fragestellung analytisch zu begegnen, weil die Möglichkeiten der Annahmenverletzung vielfältig sind. Deshalb wird hier mit Hilfe von Simulation (Kolonko, 2008) das Verhalten von IPW unter verschiedenen Graden der Annahmenverletzung untersucht. Die Schätzung durch IPW wird mit CCA verglichen, wenn die Annahme (ii) hinsichtlich der Schranke δ für die Beobachtungswahrscheinlichkeiten variiert und bezüglich $P(s_i = 1|w_i) = P(s_i = 1|v_i)$ verletzt wird. Ziel ist es, Hinweise für Datenanalysten zu erarbeiten, wann der Einsatz von IPW sinnvoll ist. Drei Hypothesen aus der Literatur werden gezielt untersucht, Zusammenhänge zwischen Gütekriterien und Kenngrößen der Schätzung werden analysiert und Hinweise aus der Literatur zum Einsatz von IPW werden zusammengestellt. Die zu untersuchenden Hypothesen sind die folgenden.

(H1) IPW reduziert die Verzerrung und verringert die Effizienz der Schätzung (Wooldridge, 2007).

(H2) Die Verwendung der geschätzten Beobachtungswahrscheinlichkeiten $\hat{\pi}_i$ führt häufiger zu korrekten Ergebnissen, als die Verwendung der wahren Beobachtungswahrscheinlichkeiten π_i. Diese Hypothese erscheint nicht intuitiv, da in diesem Fall vorhandene Informationen ignoriert werden. Die Literatur gibt aber wiederholt Hinweise für diese Hypothese (Wooldridge, 2002; Kang & Schafer, 2007; Robins, Sued, Lei-Gomez & Rotnitzky, 2007; Freedman & Berk, 2008).

(H3) Bei der Schätzung der Beobachtungswahrscheinlichkeiten wird eine Linkfunktion festgelegt, wie nachfolgend für das Selektionsmodell (12) beschrieben. Die gewählte Linkfunktion g lässt sich mit der von Hinkley (1985) vorgeschlagenen Methode verifizieren.

Für die Untersuchungen werden künstliche Stichproben simuliert. Auf Basis der Simulationen werden die Schätzungen von CCA und von IPW verglichen. Beide Methoden werden einen interessierenden Parameter a schätzen, derden Zusammenhang zwischen zwei Variablen X und Y darstellt. Die Schätzung durch CCA ist $\hat{a}_{CCA}$ und durch IPW ist $\hat{a}_{IPW}$. Ziel ist es, Hinweise zu sammeln, wann der Einsatz von IPW zu empfehlen ist. Dies wäre der Fall, wenn die Schätzung von IPW näher an a liegt, als die von CCA und wenn die Varianz der Schätzung durch IPW geringer ist (IPW eine höhere Effizizenz aufweist). Dazu werden die folgenden Gütekriterien betrachtet.

(Q1) Relative Reduktion der Verzerrung (bias reduction)

$$BR = \frac{Bias_{CCA} - Bias_{IPW}}{SE_{IPW}} \tag{4}$$

Hierbei ist $Bias_{CCA} = |a - \hat{a}_{CCA}|$, $Bias_{IPW} = |a - \hat{a}_{IPW}|$ und $SE_{IPW} = +\sqrt{Var(\hat{a}_{IPW})}$.
Positive Werte von BR deuten darauf hin, dass IPW Verzerrung reduziert hat, während negative Werte eine Vergrößerung der Verzerrung repräsentieren.

(Q2) Verhältnis der Varianzen von IPW Schätzung und CCA Schätzung (quotient of variances)

$$QV = \frac{Var(\hat{b}_{IPW})}{Var(\hat{b}_{CCA})} \tag{5}$$

Werte über 1 für QV deuten darauf hin, dass die IPW Schätzung ineffizienter ist als die Schätzung durch CCA.

(Q3) Abdeckungsrate (coverage) von a durch 95%-Konfidenzintervalle um $\hat{a}$

$$C_{CCA} = P(b \in KI_{0.95}(\hat{a}_{CCA})) \tag{6}$$

$$C_{IPW} = P(b \in KI_{0.95}(\hat{a}_{IPW})) \tag{7}$$

Es wird untersucht, ob zwischen diesen drei Gütekriterien und gewissen Kenngrößen der Schätzung Zusammenhänge existieren. Als Kenngrößen bieten sich zunächst deskriptive Variablen wie der Stichprobenumfang N und die Beobachtungsrate $N^{-1} \sum s_i$ an. Wie bereits ausgeführt, liegt die Vermutung nahe, dass die untere Grenze der Beobachtungswahrscheinlichkeiten und damit die obere Grenze der Gewichte $sup(\pi\pi_i^{-1})$ von Relevanz für die Schätzung ist. Da zu erwarten ist, dass das größte Gewicht einer starken Streuung unterworfen ist, wird zusätzlich das 0.95 - Quantil $q_{0.95}$ der Gewichte als stabilere Variable untersucht. Die Literatur empfiehlt zusätzlich den Mittelwert der Gewichte von vollständigen Beobachtungen zu betrachten (Cole & Hernán, 2008). Ebenfalls aus der Literatur (Freedman & Berk, 2008) wird bei einem korrekten Regressionsmodell empfohlen, keine Gewichtung vorzunehmen. Dementsprechend wird die Güte des Regressionsmodells mittels Determinationskoeffizient R^2_{reg} ermittelt. Zuletzt kann auch der Determinationskoeffizient R^2_{sel} für das Selektionsmodell Aufschluss über das Verhalten von IPW bieten (Kang & Schafer, 2007).

3 Simulation

Zur Untersuchung der Fragestellungen wird das Verhalten von IPW und CCA bei verschiedenen Parameterkonstellationen beobachtet. Das datenerzeugende Modell ist ein einfaches lineares Regressionsmodell (8). Zusätzlich existiert ein Selektionsmodell (10) für die Beobachtungswahrscheinlichkeiten.

$$Y = aX + bXX^o + \epsilon \tag{8}$$

$$Cov(X, X^o) = c \tag{9}$$

$$\pi_i = P(s = 1|Y, X, X^o) = f^{-1}(dX + eY + \epsilon_\pi) \tag{10}$$

Hierbei stellt Y die abhängige Variable, X und X^o die unabhängigen Variablen dar. Der inhaltlich interessierende Parameter a, der den Effekt von X auf Y darstellt, wird untersucht. Die Variable X wird vom Datenanalysten erhoben; X^o jedoch nicht ($X^{omitted}$ kurz: X^o). Aufgrund eines Interaktionseffektes von XX^o kann es bei der Schätzung von a zu Verzerrungen kommen, da X^o nicht erhoben wurde (omitted variable bias). Es gilt $X \sim \mathcal{N}(0, 1)$ und $X^o \sim \mathcal{N}(0, 1)$. Die Fehler ϵ und ϵ_π sind für jede Untersuchungseinheit unabhängig und identisch verteilt mit $\epsilon \sim \mathcal{N}(0, 1)$ und $\epsilon_\pi \sim \mathcal{N}(E(\epsilon_\pi), 1)$. Das Selektionsmodell führt eine Verzerrung aufgrund systematisch fehlender Beobachtungen für Y ein (nonresponse bias) (Bortz & Döring, 2006). Diese Verzerrung versucht IPW zu korrigieren. Die Beobachtungsrate $N^{-1} \sum s_i$ wird durch das Selektionsmodell (10) variiert; die erwartete Beobachtungsrate durch $E(\epsilon_\pi)$ manipuliert. Sei zum Beispiel die Beobachtungsrate 0.7, so wird y_i im Mittel bei 30% der Untersuchungseinheiten als unbeobachtet markiert ($s_i = 0$). Als Linkfunktion f wird für alle Simulationen die Logitfunktion verwendet $f(x) = ln(x[1 - x]^{-1})$, sodass $f^{-1}(x) = [1 + exp(-x)]^{-1}$ ist (Aldrich & Nelson, 1984).

Jede Simulation lässt sich als ein Datensatz $w_i = (s_i, y_i, x_i, x_i^o)$ mit $i = 1..N$ beschreiben, wobei y_i nicht beobachtet wird, falls $s_i = 0$ und x_i^o niemals beobachtet wird. Tabelle 1 stellt ein Schema für die simulierten Datensätze dar, wobei grau hinterlegte Zellen fehlende Daten markiert.

Die Parameter a, b, c, d, e, die Linkfunktion f aus Gleichungen 8 und 10 sowie N werden variiert. Für jede dieser Parameterkonstellation werden 100 Simulationen durchgeführt. In jeder Simulation werden N Untersuchungseinheiten zufällig erzeugt. Anhand dieser Simulationen werden die Schätzungen von CCA und IPW für a verglichen. Es werden jeweils Stichproben für $N \in \{30, 100, 200\}$ gezogen. Der Erwartungswert der Nichtbeobachtungsrate wurde durch den Fehlerterm e_π auf 10%, 20% und 30% gesetzt.

i	s_i	y_i	x_i	x_i^o
1	1			
2	0			
3	0			
4	1			
...				
N				

Tabelle 1: *Datenmodell der Simulationen. Graue Hinterlegung repräsentiert fehlende Daten.*

Fälle mit weniger vollständigen Beobachtungen werden hier nicht diskutiert, da davon ausgegangen werden sollte, dass die Analyse besonderer Vorsicht bedarf und die hier erarbeiteten, heuristischen Empfehlungen nicht anzuwenden sind. Der Haupteffekt $a \in \{0, 3, 6\}$ nimmt keine negativen Werte an, weil sich derartige Parameterkonstellationen analog zu den Parameterkonstellationen verhalten, für die das Vorzeichen der anderen Parameter vertauscht ist. Der Interaktionseffekt variiert zwischen $b \in \{-3, 0, 3\}$. Die Variablen X und X^o nehmen $c = Cov(X, X^o) = r_{XX^o} \in \{-0.7, 0.3, 0, 0.3, 0.7\}$ an (wegen $Var(X) = Var(X^o) = 1$). Analog werden Haupteffekte der Variablen X und Y auf die Beobachtungswahrscheinlichkeit von Y variiert mit $d \in \{-6, -3, 0, 3, 6\}$ und $e \in \{-3, 0, 3\}$. Dieser Simulationsrahmen deckt ist in mehreren Dimensionen beschränkt, z.B. ist unter MAR das Selektionsmodell immer korrekt spezifiziert. Die Auswirkungen dieser Einschränkungen werden in Abschnitt 5 betrachtet.

Bei der Schätzung des Koeffizienten a aus Gleichung (8) muss der Datenanalyst auf das Regressionsmodell in Gleichung (11) und das Selektionsmodell in Gleichung (12) zurückgegriffen werden, weil X^o nicht erhoben wurde. Die Konstanten α und α_π (intercept) dienen lediglich zur Schätzung und werden im Folgenden keine inhaltliche Relevanz haben.

$$\hat{Y} = aX + \alpha \tag{11}$$

$$\hat{\pi} = \hat{P}(S = 1|X) = g(eX + \alpha_\pi) \tag{12}$$

Wird auf diese Weise a geschätzt, so entsteht Verzerrung erstens durch fehlende Parameter (omitted variable bias) und durch Selektion (selection bias). Nach Hypothese H1 soll bei der Simulation die Verzerrungsreduktion BR positiv werden und der Quotient der Varianzen QV über 1 liegen.

Zur Untersuchung der Hypothese H2 wird die Verwendung der wahren Beobachtungswahrscheinlichkeiten aus dem datenerzeugenden Modell (10) verglichen mit der Verwendung von geschätzten Beobachtungswahrscheinlichkeiten. Im letzteren Fall dient das Selektionsmodell (12) zur Schätzung der Beobachtungswahrscheinlichkeiten $\hat{P}(s_i = 1|x_i) = \hat{\pi}_i$ und damit zur Berechnung der Gewichte $\hat{\pi}\hat{\pi}_i^{-1}$ aus Gleichung (3). Um Hypothese H3 zu untersuchen werden für jede Simulation Selektionsmodelle (12) mit Logit, Probit und Cauchit als Linkfunktion g aufgestellt. Mit dem daraus geschätzten, linearen Prädiktoren ($\hat{\eta}_i = \hat{e}x_i$) wird ein neues Selektionsmodell aufgestellt mit $\hat{P}(S = 1|X = x) = g(\hat{\eta} + \hat{h}\hat{\eta}^2)$. Hinkley (1985) empfiehlt die Interpretation, dass die gewählte Linkfunktion g ungleich der wahren Link Funktion f ist, wenn $\hat{h}$ signifikant von 0 verschieden ist. Bei der Datengeneration der gesamten Simulation wurde für f die Logitfunktion verwendet.

Die Simulation erfolgt mit Hilfe der freien statistischen Software R (R Core Team, 2012). Insbesondere zur Generierung der Gewichte von IPW wird das Softwarepaket ipw Version 2.15.3 von Wal und Geskus (2011) genutzt. Zur Untersuchung der Linkfunktionen wurde das Paket VGAM Version 2.15.3 verwendet.

Beispielsimulation

Eine Simulation sei zur Veranschaulichung kurz charakterisiert. Beispielsweise wird die Parameterkonstellation $a = 6$, $b = 3$, $c = -0.7$, $d = -3$ und $e = 0$ (MAR) betrachtet. Der Stichprobenumfang beträgt $N = 200$. Die verwendete Linkfunktion in (10) ist die Logitfunktion. 128 der Untersuchungseinheiten wurden vollständig beobachtet. Ein Streudiagramm für die generierten Daten sowie durch CCA und IPW geschätzte Regressionsgeraden sind in Abbildung 1 zu sehen. Eine geschätzte Verteilungsfunktion der Gewichte lässt sich in Abbildung 2 nachvollziehen. Es treten Gewichte auf, die einzelnen Untersuchungseinheiten einen größeren Einfluss auf die Schätzung verleihen. In diesem Fall erhalten die Beobachtungen mit $X > 1$ große Gewichte, weil die Beobachtungsrate für Untersuchungseinheiten mit $X > 1$ gering ist. So nähert sich IPW einer geschätzten Regressionsgeraden an, die geschätzt worden wäre, wenn alle Untersuchungseinheiten vollständig beobachtet gewesen wären.

Nun werden Kenngrößen und Gütekriterien für diese Simulation berechnet und über 100 Wiederholungen gemittelt. So werden Mittelwerte der Kenngrößen und Gütekriterien für alle 6075 Parameterkonstellationen gefunden.

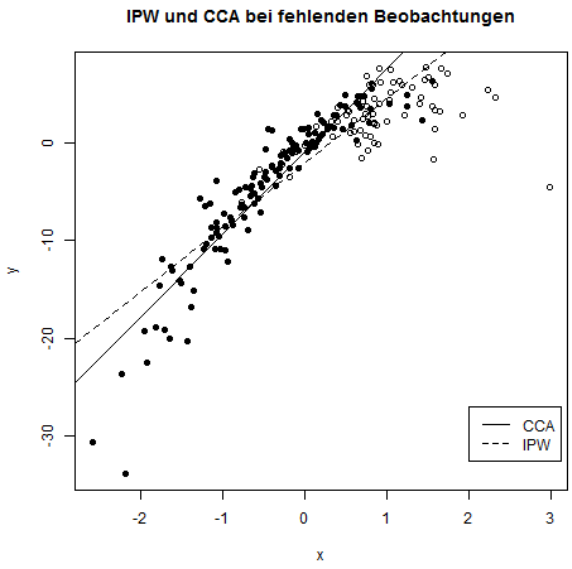

Abbildung 1: *Beispiel einer durchgeführten Simulation. Ausgefüllte Punkte repräsentieren vollständig beobachtete Untersuchungseinheiten. Kreise repräsentieren Untersuchungseinheiten, bei denen nur X beobachtet wurde.*

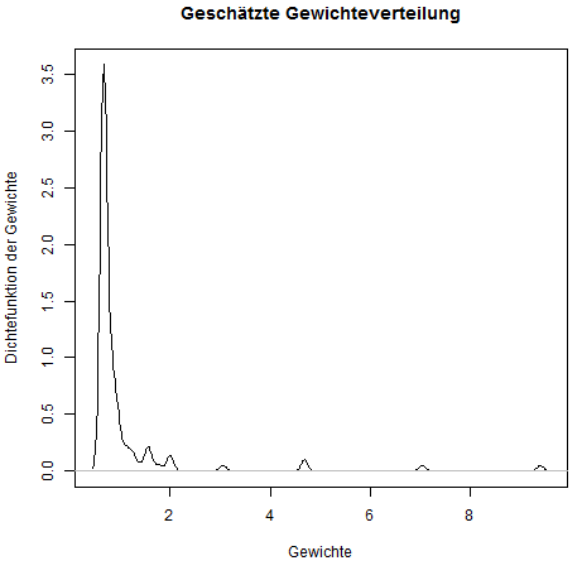

Abbildung 2: *Beispiel der geschätzten, rechtsschiefen Dichtefunktion.*

4 Ergebnisse

Um einen Eindruck über die vorliegenden Ergebnisse der Simulation zu bekommen, werden die Parameterkonstellationen in Klassen gegliedert und die Mittelwerte (mean) der Gütekriterien betrachtet sowie die empirischen Standardfehler (sd) dieser. In Tabelle 2 wird die Simulationen nach der Abhängigkeitsstruktur des Selektionsmodells (MCAR mit $d = e = 0$, MAR mit $d \neq 0$, $e = 0$ und NMAR mit $e \neq 0$) geteilt.

Für die Schätzung von $Var(\hat{a}_{IPW})$ wurden zwei Wege gewählt. Erstens wurde die Varianz durch der Methode glm() geschätzt. Zweitens wurde über die 100 Simulationen einer Parameterkonstellation die Varianz von $\hat{a}$ empirisch geschätzt. Ein Vergleich dieser beiden Methoden ergibt, dass die Varianzschätzung durch glm() fälschlicherweise für gewichtete Methoden sogar kleinere Varianzen ausgibt. Dies kann aus analytischen Gründen abgelehnt werden (Wooldridge, 2007). Die empirische Varianzschätzung von $\hat{a}$ ergibt hingegen $mean(QV) = 4.326$. Dieses Ergebnis deutet darauf hin, dass die Varianz sich bei Gewichtung vervielfacht. So fanden auch Freedman und Berk (2008) sowie Seaman et al. (2012) eine Vervielfachung der Standardabweichung unter IPW. Curtis et al. (2007) empfiehlt daher die Standardabweichung mit Bootstrapmethoden zu schätzen (Efron & Tibshirani, 1993). Die Berechnung der Konfidenzintervallabdeckungen basierte auf der Varianzschätzung von glm(). So sind die berechneten Konfidenzintervalle für IPW zu klein. Deshalb wird die Abdeckungsrate C_{IPW} unterschätzt. Dementsprechend wurde dieses Kriterium aus der Untersuchung ausgeschlossen.

Es konnte festgestellt werden, dass IPW je nach Art der Annahmenverletzung Verzerrung reduzieren kann. Wenn MCAR gilt, existiert keine Verzerrung durch Selektion, die IPW reduzieren könnte ($BR \in [-0.002, 0.002]$ mit Fehlerwahrscheinlichkeit $\alpha = 0.05$). Unter MAR kann IPW die Verzerrung reduzieren ($BR \in [0.303, 0.344]$). Unter NMAR erhöht IPW die Verzerrung ($BR \in [-0.505, -0.47]$). Insgesamt erhöht IPW die Varianz in 5092 von 6075 Parameterkonstellationen signifikant (83.8%).

	MCAR	MAR	NMAR	Total
mean(BR)	0	0.324	-0.488	-0.239
sd(BR)	0.017	0.415	0.562	0.621
mean(QV)	2.55	4.65	4.374	4.326

Tabelle 2: *Mittelwerte und Standardabweichungen der Gütekriterien unterteilt in Klassen der Abhängigkeitsstruktur von Nichtbeobachtung.*

Es lässt sich festhalten, dass IPW im überwiegenden Anteil der Anwendungen die Varianz der Schätzung erhöht. Somit lässt sich die Hypothese H1 bestätigen, da IPW unter korrekten Annahmen (i) und (ii) die Verzerrung reduziert und unabhängig der Annahmenverletzung die Effizienz verringert (Varianz der Schätzung erhöht). Allerdings ist das positive Abschneiden bezüglich der Verzerrungsreduktion von IPW unter MAR kritisch zu betrachten, weil das Selektionsmodell für alle Simulationen unter MAR korrekt spezifiziert wurde.

Die Verwendung geschätzter Beobachtungswahrscheinlichkeiten liefert effizientere Ergebnisse als die Verwendung der wahren Beobachtungswahrscheinlichkeiten. Beim Vergleich der relativen Verzerrungsreduktion BR in der gesamten Simulation ergibt der t-Test einen signifikanten Unterschied zugunsten der geschätzten Beobachtungswahrscheinlichkeiten ($t = 9.28, p < 0.5$). Die Verwendung geschätzter Beobachtungswahrscheinlichkeiten scheint in Anbetracht der relativen Verzerrungsreduktion empfohlen zu sein, sodass auch Hypothese H2 hier Bestätigung findet.

Bei der Prüfung der Linkfunktion wird die Logitfunktion im Mittel in 16 von 100 Simulationen fälschlicherweise als Link-Funktion abgelehnt. Die Cauchitfunktion wurde in 20 von 100 und die Probitfunktion in 16 von 100 Simulationen korrekterweise abgelehnt. Die Unterschiede zwischen den Funktionen scheinen sich hier aus dem Anstieg der jeweiligen Funktion zu ergeben (Kang & Schafer, 2007). Da die gewählten Linkfunktionen eine ähnliche Ablehnungsrate aufweisen, lässt sich kein Hinweis darauf finden, dass die Linkfunktion des Selektionsmodells mit der beschriebenen Methode verifiziert werden kann. Diese Aussage gilt jedoch nur für den eingeschränkten Variationsbereich der Parameter und wenn die datenerzeugende Funktion eine Logitfunktion ist. Die Hypothese H3 lässt sich somit in dieser Simulation nicht bestätigen, jedoch bleibt offen, ob ein erweiterten Simulationsrahmen andere Ergebnisse liefert.

Im Folgenden sollen praktische Implikationen aus der vorliegenden Simulationsstudie gezogen werden. Die Simulationen wurden daraufhin untersucht, ob Zusammenhänge zwischen der relativen Verzerrungsreduktion BR und den Kenngrößen existieren, die dem Datenanalysten zugänglich sind. So wäre es dem Datenanalysten möglich, anhand der ihm bekannten Kenngrößen abzuschätzen, ob der Einsatz von IPW empfohlen ist. Unter MCAR ist die Verwendung von IPW nicht zu empfehlen. Diese Fälle lassen sich durch einen kleinen Determinationskoeffizient identifizieren ($R^2_{sel} < 0.1$).

Der Determinationskoeffizienten kann allein durch Hinzunahme selbst unabhängiger Variablen größer werden, weshalb hier der korrigierte Determinationskoeffizient betrachtet werden sollte. Unter MAR ergibt eine Zusammenhangsanalyse von BR auf die Kenngrößen, dass die Verzerrung mit Hilfe von IPW sinkt, je größer R^2_{reg}, je größer N und je kleiner die Abdeckungsrate $N^{-1} \sum s_i$ ist. Freedman und Berk (2008) empfiehlt, bei korrekt spezifizierten Regressionsmodellen keine Gewichtung vorzunehmen. Der Hinweis von Cole und Hernán (2008) bestätigt sich auch in dieser Untersuchung, wonach der Mittelwert der Gewichte für vollständig beobachtete Untersuchungseinheiten nicht signifikant größer als 1 sein darf. Zu erwähnen ist, dass über die gesamte Untersuchung die Kenngrößen $sup(\pi\pi_i^{-1})$ und $q_{0.95}$ keinen zusätzlichen Zusammenhang aufklären. Der Einfluss der größten Gewichte scheint durch den Mittelwert der Gewichte moderiert zu werden.

Die beschriebenen Heuristiken lassen wie folgt zusammenfassen. Das Selektionsmodell sollte einen signifikant von 0 verschiedenen Determinationskoeffizienten besitzen. Das Regressionsmodell sollte einen möglichst großen Regressionskoeffizienten aufweisen (mindestens 0.3), jedoch nur wenn weiterhin von einer Fehlspezifikation auszugehen ist. Der Mittelwert der Gewichte darf 1 nicht überschreiten. Der Stichprobenumfang sollte groß genug sein um trotz Nichtbeobachtungsrate von bis zu 30% genügend vollständige Beobachtungen aufzuweisen ($N > 100$).

Zusammenfassend ist die Art der Abhängigkeitsstruktur der Nichtbeobachtung (MCAR, MAR, NMAR) entscheidend für die Leistung von IPW. Ist Verzerrung nicht auf Selektion begründet (MCAR), erhöht IPW lediglich die Varianz der Schätzung. Wird die Selektion falsch geschätzt, was unter NMAR möglich ist, kann es zu einer Vergrößerung der Verzerrung kommen. Dagegen stellen sich hier die konkrete Ausprägung der unteren Schranke δ für die Beobachtungswahrscheinlichkeiten nur im Kontext der anderen Gewichte als geeignet heraus, um die Leistung von IPW vorherzusagen.

5 Diskussion

Wenn systematische Nichtbeobachtung in einer ansonsten repräsentativen Stichprobe vorliegt, kann neben anderen Alternativen IPW eingesetzt werden. IPW reduziert nur unter bestimmten Bedingungen die Verzerrung von Schätzern. Die Effizienz der Schätzung wird jedoch in jedem Fall dabei verringert, sodass generell größere Stichproben benötigt werden, um präzise Aussagen treffen zu können. Deshalb sollten die Kenngrößen der Berechnung zur Entscheidung herangezogen werden, ob das Ergebnis der IPW Schätzung haltbar ist. Bei der Anwendung von IPW sollten in jedem Fall die Beobachtungswahrscheinlichkeiten für die Untersuchungseinheiten geschätzt werden, selbst wenn die wahren Beobachtungswahrscheinlichkeiten bekannt sind. Dieses nicht intuitive Ergebnis liegt darin begründet, dass Selektion im Kontext einer Stichprobe gesehen werden muss. Die Nichtbeobachtung in einer Stichprobe bietet bessere Hinweise darauf, wie Verzerrung (in dieser Stichprobe) zu reduzieren ist, als die wahren Beobachtungswahrscheinlichkeiten. In dieser Untersuchung konnte nicht bestätigt werden, dass die gewählte Linkfunktion für das Selektionsmodell verifiziert oder falsifiziert werden kann.

Dass die Methode zur Verifizierung der Linkfunktion keine Ergebnisse lieferte, ist möglicherweise darin begründet, dass das Selektionsmodell nicht ausreichend variiert wurde. Der Untersuchungsrahmen ist dafür möglicherweise zu klein gewählt. Eine Hinzunahme von quadratischen Termen in das datenerzeugende Selektionsmodell ergibt möglicherweise andere Ergebnisse hinsichtlich der Überprüfung von Linkfunktionen. Weiterhin führt die Einschränkung auf nur zwei abhängige Variablen im datenerzeugenden Selektionsmodell möglicherweise auch zu einer Überschätzung der Leistung von IPW in manchen Untersuchungsabschnitten. Die Ergebnisse eines größeren Untersuchungsrahmens hätten jedoch den Raum dieser Bachelorarbeit überschritten.

Als Ergebnis dieser Untersuchung ist festzuhalten, dass IPW nur dann eingesetzt werden sollte, wenn sowohl Selektions- als auch Regressionsmodell jeweils korrigierte Determinationskoeffizienten $R^2 \geq 0.3$ besitzen. Dies setzt voraus, dass immer beobachtbare Variablen identifizierbar sind, die zur Schätzung der Nichtbeobachtungswahrscheinlichkeiten ausreichen. Die berechneten Gewichte sollten in so fern moderat sein, als dass der Mittelwert der Gewichte $mean(\pi\pi_i^{-1}) \leq 1$ betragen soll. Ferner ist der Einsatz ist IPW ein größeren Stichprobenumfang ($N > 100$) mit entsprechend angemessener Beobachtungsrate $N^{-1} \sum s_i \geq 0.7$ zu empfehlen. Unter diesen Umständen hilft IPW die Verzerrung der Schätzung zu reduzieren.

Weitere Untersuchungen wären notwendig, um zu prüfen, ob die Leistung von IPW tatsächlich so positiv ausfällt. Dazu sollten Modelle mit mehr Variablen hinzugezogen und deren Haupteffekte variiert werden, um höhere Grade der Fehlspezifikationen des Regressions- und Selektionsmodells zu erlauben. Ebenfalls sollte die Verteilung der Fehler, die hier einheitlich als normalverteilt gewählt wurden, variiert werden. Es lässt sich vermuten, dass die Abhängigkeit der Nichtbeobachtung von fehlenden Variablen umso weniger schädlich ist, je größer der Zusammenhang zwischen diesen fehlenden Variablen und immer beobachteten Variablen ist. Ein möglicher Untersuchungsansatz besteht darin, das datenerzeugende Modell um eine unabhängige Variable zu erweitern, die nicht immer beobachtet wird, aber dennoch einen Einfluss auf die Beobachtungswahrscheinlichkeiten hat. Die zu prüfende Hypothese ist, ob dann mit steigendem Zusammenhang der unabhängigen Variablen, die Leistung von IPW besser wird. Ein weiterer Untersuchungsansatz gilt den Maßnahmen, die Gewichte anzupassen. Dazu gehört Truncation und Diskretisierung der unabhängigen Variablen des Selektionsmodells. Die Auswirkungen dieser Methoden sollten im Licht der Beziehung zwischen Effizienz und Verzerrungsreduktion untersucht werden.

6 Literaturverzeichnis

Aldrich, J. H. & Nelson, F. D. (1984). *Linear probability, logit, and probit models* (Bd. 45). Sage.

Bortz, J. & Döring, N. (2006). *Forschungsmethoden und evaluation: für Human-und Sozialwissenschaftler.* Springer DE.

Cole, S. R. & Hernán, M. A. (2008). Constructing inverse probability weights for marginal structural models. *American journal of epidemiology*, *168*(6), 656–664.

Curtis, L. H., Hammill, B. G., Eisenstein, E. L., Kramer, J. M. & Anstrom, K. J. (2007). Using inverse probability-weighted estimators in comparative effectiveness analyses with observational databases. *Medical care*, *45*(10), S103–S107.

Efron, B. & Tibshirani, R. (1993). *An introduction to the bootstrap* (Bd. 57). CRC press.

Freedman, D. A. & Berk, R. A. (2008). Weighting regressions by propensity scores. *Evaluation Review*, *32*(4), 392–409.

Haapea, M., Veijola, J., Tanskanen, P., Jääskeläinen, E., Isohanni, M. & Miettunen, J. (2011). Use of inverse probability weighting to adjust for non-participation in estimating brain volumes in schizophrenia patients. *Psychiatry Research: Neuroimaging*, *194*(3), 326–332.

Hinkley, D. (1985). Transformation diagnostics for linear models. *Biometrika*, *72*(3), 487–496.

Kang, J. D. & Schafer, J. L. (2007). Demystifying double robustness: A comparison of alternative strategies for estimating a population mean from incomplete data. *Statistical science*, *22*(4), 523–539.

Kolonko, M. (2008). *Stochastische Simulation: Grundlagen, Algorithmen und Anwendungen.* Springer DE.

McClintock, P. V. (2010). Missing Data in Longitudinal Studies: Strategies for Bayesian Modeling and Sensitivity Analysis, by MJ Daniels and JW Hogan: Scope: research monograph. Level: professional statisticians and Ph. D. students.

R Core Team. (2012). *R: A Language and Environment for Statistical Computing.* Vienna, Austria. (ISBN 3-900051-07-0)

Robins, J., Sued, M., Lei-Gomez, Q. & Rotnitzky, A. (2007). Comment: Performance of Double-Robust Estimators WhenÏnverse Probability"Weights Are Highly Variable. *Statistical Science*, *22*(4), 544–559.

Schafer, J. L. & Graham, J. W. (2002). Missing data: our view of the state of the art. *Psychological methods*, *7*(2), 147.

Seaman, S. R., White, I. R., Copas, A. J. & Li, L. (2012). Combining Multiple Imputation and Inverse-Probability Weighting. *Biometrics*, *68*(1), 129–137.

Wal, W. M. van der & Geskus, R. B. (2011). Ipw: an R package for inverse probability weighting. *Journal of Statistical Software*, *43*(i13).

Wooldridge, J. M. (2002). Inverse probability weighted M-estimators for sample selection, attrition, and stratification. *Portuguese Economic Journal*, *1*(2), 117–139.

Wooldridge, J. M. (2007). Inverse probability weighted estimation for general missing data problems. *Journal of Econometrics*, *141*(2), 1281–1301.

Selbstständigkeitserklärung

Ich versichere, dass ich die Bachelorarbeit im Studiengang Mensch-Computer-Interaktion selbstständig verfasst und keine anderen als die angegebenen Hilfsmittel - insbesondere keine im Quellenverzeichnis nicht benannten Internet-Quellen - benutzt habe. Alle Stellen, die wörtlich oder sinngemäß aus Veröffentlichungen entnommen wurden, sind als solche kenntlich gemacht. Ich versichere weiterhin, dass ich die Arbeit vorher nicht in einem anderen Prüfungsverfahren eingereicht habe und die eingereichte schriftliche Fassung der auf dem elektronischen Speichermedium entspricht.

Mit der Einstellung der Arbeit in die Bibliothek bin ich einverstanden.