EBERHARD KARLS
UNIVERSITÄT
TÜBINGEN

Steffen Hillmert

Die Messung beruflicher Konsequenzen des Studiums: Ein simulationsbasierter Vergleich von Regression und Propensity-Score-Matching als Verfahren der Kausalanalyse

ESOC Working Paper 2/2010

Steffen Hillmert

Institut für Soziologie
Universität Tübingen
Wilhelmstr. 36
72074 Tübingen
steffen.hillmert@uni-tuebingen.de

Zusammenfassung

Dieses Papier vergleicht exemplarisch zwei verschiedene Methoden zur Kontrolle von Drittvariablen bei kausalorientierten Analysen: multivariate Regression und Matching mittels Propensity Scores. Diskutiert werden die Grundannahmen und Implikationen der unterschiedlichen Verfahren. Um die Rahmenbedingungen möglichst einheitlich zu gestalten, werden die Verfahren mittels einer einfachen Simulationsstudie verglichen. Das Hochschulstudium gilt in diesem Sinne als „Treatment“, der berufliche Status in der Folgezeit als Messgröße. Implikationen der Verfahrensunterschiede werden erläutert.

Steffen Hillmert

Die Messung beruflicher Konsequenzen eines Studiums:
Ein simulationsbasierter Vergleich von Regression und Propensity-Score-Matching als Verfahren der Kausalanalyse

1. Beispiel „Bildungsrenditen“

Dieses Papier vergleicht exemplarisch verschiedene Methoden zur Kontrolle von Drittvariablen, die zu unterschiedlichen Größenordnungen geschätzter Effekte führen können. Der Versuch der Bestimmung von Kausaleffekten erfolgt stets vor dem Hintergrund bestimmter Modellvorstellungen und Annahmen, die es zu berücksichtigen gilt. Es gibt offensichtlich nicht *das* in jedem Fall korrekte Verfahren, doch lassen sich möglicherweise Rahmenbedingungen spezifizieren, unter denen das eine oder das andere Verfahren angemessener ist.

Obwohl somit der methodische Aspekt im Vordergrund steht, ist dieser nicht unabhängig von den Inhalten. Diesbezüglich steht die Frage nach Bildungseffekten vor dem Hintergrund einer möglichen (weiteren) Expansion des Bildungssystems im Fokus. Praktisch alle Industriegesellschaften haben in der zweiten Hälfte des 20. Jahrhunderts eine größere Bildungsexpansion erlebt, welche die allgemeine, aber auch die berufliche und die akademische Ausbildung betraf (zu den Konsequenzen für Ungleichheitsrelationen und andere Konsequenzen vgl. Müller 1998; Becker & Hadjar 2006). Zudem gibt es seit Jahren eine Diskussion um eine Weiterführung der Bildungsexpansion, die sich u.a. um die Frage dreht, ob die Hochschulen weiter ausgebaut werden sollen. Diskutiert wird dies zumeist vor dem Hintergrund der aktuellen demografischen Entwicklung, die in vielen Bereichen langfristig einen Fachkräftemangel erwarten lässt, bzw. unter dem Schlagwort einer generell notwendigen Höherqualifizierung der Erwerbsbevölkerung im Zuge der Herausbildung der „Wissensgesellschaft“. Hingewiesen wird allerdings oft auch auf *individuelle* Effekte, insbesondere in dem Sinn, dass sich „ein Studium (individuell) lohnt...“. Nur dieser Themenkomplex soll im Folgenden betrachtet werden.

Zunächst ist ein solcher Nutzen zu spezifizieren (in formaler am bekanntesten ist diesbezüglich wohl die Humankapitaltheorie in der Tradition von Becker 1964; Mincer 1974).

Mögliche Dimensionen sind hier finanzielle Effekte, Schutz vor Arbeitslosigkeit aber auch nicht-materielle bzw. nicht arbeitsbezogene Konsequenzen. In diesem Papier soll beruflicher Status (Berufsprestige) als Beispiel dienen. Dabei stehen aber methodische Fragen im Vordergrund: Wie und mit welchen Verfahren lassen sich solche Effekte messen? Welche Schlussfolgerungen lassen sich aus den Ergebnissen ziehen?

Verschiedene Studien haben Zusammenhänge zwischen Studienerfahrungen und dem anschließenden Erwerbsverlauf gefunden. Im Folgenden wird unter der Messung von Kausaleffekten nicht nur die Frage verstanden, *ob* es Einflüsse des Studiums auf Erwerbsverlauf gibt (und welche Richtung sie haben), sondern auch, *wie groß* sie sind. Statistische Gruppendifferenzen – wie jene im mittleren beruflichen Status zwischen Studierten und nicht Studierten – sind allerdings zunächst nur deskriptiv. Die Interpretation dieser Unterschiede als *Effekte* bzw. *Kausaleffekte* des Studiums setzt voraus, dass diese Unterschiede tatsächlich auf das Studium und nicht etwa auf andere Faktoren zurückgehen. Dies betrifft zunächst das konventionelle Standardverfahren, die multivariate Regressionsanalyse. Neuere Diskussionen in der Statistik („kontrafaktische Modelle"/Rubins Kausalmodell) können darüber hinaus systematische Orientierung bieten. Sie legen ein quasi-experimentelles Untersuchungsdesign zugrunde und werden praktisch etwa in der Form von Matchingverfahren realisiert (vgl. Gangl/DiPrete 2004). Mit der Konzentration auf die Auswirkungen eines eindeutigen Faktors („Treatment" im Sinne eines Experiments) ist der Anwendungsbereich dieser Verfahren allerdings auf den Vergleich weniger Untersuchungsgruppen beschränkt; nicht geeignet hierfür ist etwa die Frage nach dem Einfluss kontinuierlicher Variablen. In diesem Sinne eignet sich die Bestimmung von Effekten eines Ereignisses wie das Vorhandensein eines Studiums („ja/nein") prinzipiell für die Anwendung solcher Verfahren[1].

2. Messprobleme und Verfahren der Kausalanalyse

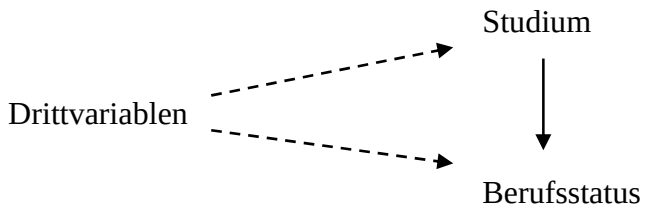
2.1 Drittvariablen im Regressionsmodell

Deskriptiv ist die Frage nach Statusunterschieden zwischen Studierten und Nicht-Studierten relativ einfach zu beantworten. Die Messung und Interpretation von *Effekten* ist hingegen viel schwieriger. Das Grundproblem besteht in möglichen Drittvariableneffekten. Gehen also

[1] Eher problematisch an diesem Beispiel scheint hingegen zu sein, das Bildungsprozesse nicht denkbar ohne aktive Beteiligung (und häufig langfristige Planungen) der betreffenden Individuen, die Idee von mehr oder weniger zwangsläufigen Effekten einer „Behandlung" also eine sehr grobe Vereinfachung darstellt.

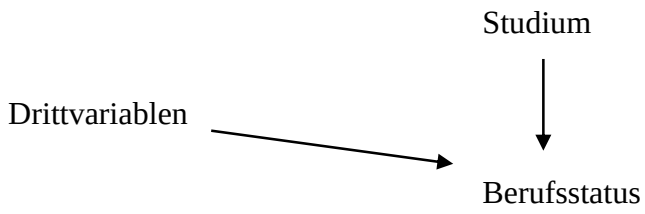
beobachtete Unterschiede möglicherweise gar nicht auf das Studium, sondern auf Faktoren zurück, die bereits davor vorlagen? Genauer gesagt können solche Drittvariablen (a) auf die Wahrscheinlichkeit, zu studieren und /oder (b) auf die Messgröße Berufsstatus wirken und so die Schätzung des Kausaleffekts des Studiums verzerren (vgl. Abbildung 1). Die hier verglichenen Verfahren versuchen diese Einflüsse jeweils an unterschiedlicher Stelle zu kontrollieren.

Abbildung 1: Vereinfachte Kausalmodell der Auswirkung des Studiums auf den Berufsstatus



Das klassische Verfahren zur Bestimmung von Effekten ausgewählter Variablen im Rahmen von Beobachtungsstudien ist die multivariate Regressionsanalyse, bei der eine mehr oder minder große Anzahl von Drittvariablen hinsichtlich ihrer Bedeutung für die abhängige Variable „kontrolliert“ wird und somit der verbleibende „Partialeffekt“ der interessierenden Variable – im vorliegenden Fall des Studiums – bestimmt wird (vgl. Abbildung 2).

Abbildung 2: Kausalmodell für eine „Kontrolle“ von Drittvariablen im Regressionsmodell



2.2 (Quasi-)experimentelle Kausalanalyse

Einschlägig für kausalorientierte Studien ist aber eher das Experimentaldesign, das sich u.a. auszeichnet durch einen kontrollierten Gruppenzugang zu mindestens einer Untersuchungs- und einer Kontrollgruppe, eine gruppenspezifische Behandlung (*Treatment*) und eine nachgelagerte Messung der interessierenden Variable. Es gibt eine längere Diskussion darüber, ob auch Beobachtungsstudien prinzipiell die Basis von Kausalaussagen vergleichbarer Qualität bilden können. Eine gewisse Ernüchterung folgt aus den Ergebnissen von LaLonde (1986), der Experimentaldaten wie Beobachtungsdaten behandelte und damit trotz fortgeschrittener Analyseverfahren (mit Kontrollvarieblen) alle möglichen verzerrten Ergebnisse erhielt. Das Experimentaldenken hat seither auch in den Sozialwissenschaften wieder an Bedeutung gewonnen, und aus der Statistik sind systematisierende theoretische Überlegungen gekommen, welche versuchen, die (quasi-)experimentelle Logik auch für Beobachtungsdaten nutzbar zu machen. Dies betrifft insbesondere die „kontrafaktischen" Kausalitätsüberlegungen im Anschluss an die Arbeiten von Donald Rubin („Rubins Kausalmodell") (Rubin 1974; Morgan/Winship 2007; Longford 2008). Die Nähe zum Experimentaldesign spiegelt sich nicht zuletzt in den verwendeten Begriffen wider, die der im Zusammenhang mit Experimenten üblichen Terminologie entlehnt sind. So spricht man auch hier üblicherweise vom *Treatment* als der Variable, deren kausaler Effekt bestimmt werden soll – auch wenn es sich nicht um eine kontrollierte „Behandlung" im eigentlichen Sinne handelt.

Zentral im Experiment ist die *zufällige Zuordnung* zu Treatment- und Kontrollgruppe. Dies bedeutet insbesondere auch, dass Selbstselektion, wie sie bei der selbständigen Wahl der jeweiligen Gruppe durch die Teilnehmer auftritt, unterbunden wird. Im sozialen Leben sind Selbstselektionen oder andere Formen der systematisch verzerrten Gruppenzuordnung die Regel; eine zufällige Zuordnung ist eigentlich nur im Experimentaldesign zu verwirklichen. Dennoch kann diese Eigenschaft des Experimentaldesigns als sinnvolle Referenz auch für Beobachtungsstudien gelten. Hier setzen die neueren Überlegungen zur Kausalanalyse an.

Was die Forschenden i.d.R. eigentlich interessiert der *individuelle kausale Effekt* eines Treatments (*Individual Causal Effect* oder kurz ICE) bei jedem Individuum i, formal bestimmbar als Differenz d: $d_i = Y_i\ (t) - Y_i\ (c) = Y$ (unter ‚Treatment') – Y (unter ‚Kontrollbedingung').

Dieser Effekt kann bei jedem Einzelnen unterschiedlich sein. Das Hauptproblem aber ist, dass er grundsätzlich nicht beobachtbar ist. Ein konkretes Individuum befindet sich ja stets *entweder* in der Untersuchungs- *oder* der Kontrollgruppe; insofern ist der ICE eine kontrafaktische Größe. I.d.R. greift man zu einer Vereinfachung und interessiert sich nicht für den individuellen, sondern den durchschnittlichen Kausaleffekt (*Average Causal Effect* oder kurz ACE), also die durchschnittliche Differenz zwischen den Ergebnissen unter Treatment- und Kontrollbedingungen. Allerdings ist auch der ACE kontrafaktisch, da weiterhin jedes Individuum nur in einer der beiden Situationen beobachtet werden kann. Nicht mit dem ACE zu verwechseln ist also die einfach zu beobachtende Differenz: Y (tatsächliche Treatmentgruppe) – Y (tatsächliche Kontrollgruppe), denn hierbei handelt es sich ja um zwei disjunkte Mengen handelt, nicht um den Vergleich der Situation jedes Einzelnen zwischen faktischen und kontrafaktischen Bedingungen[2].

Es gibt aber zwei *hinreichende* Bedingungen für konsistente Schätzung des ACE (Morgan/Winship 2007): (1) der Mittelwert der Treatmentgruppe unter Treatment-Bedingungen ist gleich Mittelwert Kontrollgruppe unter Treatment-Bedingungen und (2) der Mittelwert der Treatmentgruppe unter Kontroll-Bedingungen ist gleich Mittelwert Kontrollgruppe unter Kontroll-Bedingungen

Im Experimentaldesign wird dies durch Randomisierung erreicht. Genauer gesagt, gilt dies im Mittel, da weiterhin Zufallsfehler auftreten können. Auch wenn eine Randomisierung, wie in Beobachtungsstudien, nicht möglich ist, ist die Bedingung der sogenannten. *Ignorability* („Unbeachtlichkeit“) hinreichend. Dies bedeutet, dass die Ergebnisse unter beiden (potenziellen) Bedingungen unabhängig von der jeweiligen Zuordnung sind; vereinfacht gesagt ist die Zuordnung „unabhängig davon, was jeweils zu erwarten wäre“. Das bedeutet aber *nicht*, dass faktische Zuordnung und tatsächlich beobachtetes Ergebnis unabhängig sind. Dieser Zusammenhang soll ja untersucht werden! In randomisierten Experimenten ist die Bedingung der Ignorability i.d.R. gegeben, denn auch potenzielle Ergebnisse können ja als unbeobachtete Variablen aufgefasst werden, welche durch den Randomisierungsrozess (im Mittel) ausgeglichen werden.

Die beobachtete Gruppendifferenz setzt sich prinzipiell immer aus mehreren Komponenten zusammen: (1) dem wahren ACE, (2) der Baseline-Differenz und (3) der ACE-Differenz

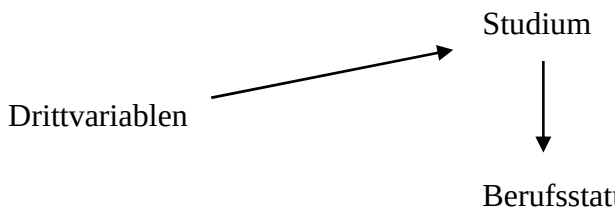
[2] In diesem Sinne liegt hier ein Problem „fehlender Daten“ vor.

zwischen den beiden Gruppen, gewichtet mit dem (inversen) Anteil der Stichprobe, der sich in der Treatmentgruppe befindet. Um den wahren ACE bestimmen zu können, müssen also die beiden anderen Komponenten eliminiert oder gemessen werden. *Baseline-Differenz:* Diese wird in multivariaten statistischen Auswertungen i.d.R. berücksichtigt und bezieht sich auf unterschiedliche Ausgangsbedingungen. Diese Selektivität kann man häufig durch Messungen und Kontrollvariablen relativ gut kontrollieren, wenngleich mögliche fehlende Variablen ein Problem darstellen. *ACE-Differenz zwischen den beiden Gruppen:* Diese wird oft nicht berücksichtigt. Hierbei handelt es sich um gruppenspezifische Effekte, die unmittelbar mit dem Treatment zusammenhängen. Das klassische Beispiel ist die Medizin, die *nur bei Kranken* das Wohlbefinden steigert. Kann man also begründet annehmen, dass der Treatment-Effekt für Angehörige von Treatment- und Kontrollgruppe gleich ist? Oft ist dies nicht der Fall, und es wird daher nur versucht, den durchschnittlichen Treatment-Effekt für die Treatmentgruppe (*Average Treatment Effect on the Treated* oder kurz ATT) zu bestimmen, also die mittlere Differenz der Ergebnisse unter Treatment- und Kontrollbedingungen *für Angehörige der Treatmentgruppe.*

2.3 Balancierung und Matching

Die im Folgenden dargestellten praktischen Verfahren wurden im engen Anschluss an die theoretischen Überlegungen der kontrafaktischen Kausalmodelle entwickelt. Genauer gesagt wurden hier Grundprinzipien wiederentdeckt, welche an sich viel einfacher und älter als das der Regressionsanalyse sind. Zentraler Ansatzpunkt ist hier die Kontrolle des Gruppenzugangs, in unserem Fall des Zugangs zu den beiden Vergleichsgruppen „Studierte“ und „Nicht-Studierte“.

Abbildung 3: Kausalmodell für eine „Kontrolle“ von Drittvariablen durch kontrollierte Gruppenbildung



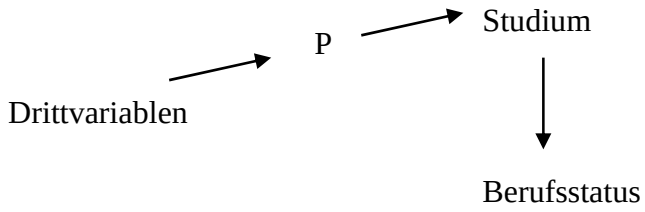
Im Experimentaldesign gibt es klassischerweise zwei Möglichkeiten, gleichartige Versuchs- und Kontrollgruppen herzustellen: die Randomisierung des Gruppenzugangs oder die Stratifizierung bzw. das Matching nach bestimmten Variablen.

Die Zuordnung zu Treatment- und Kontrollgruppe soll unabhängig von anderen Merkmalen erfolgen (T unabhängig von X). Die *gleiche Merkmalsverteilung* in beiden Gruppen bedeutet ganz offensichtlich, dass diese Unabhängigkeit gegeben ist. Bei der Randomisierung erfolgt die Zuordnung zufällig, und es kann daher eine gleiche Merkmalsverteilung in den Gruppen erwartet werden. (Genauer gesagt, kann man dies im Mittel erwarten, da weiterhin Zufallsfehler auftreten können). Eine solche Merkmalsverteilung lässt sich aber auch vorab festlegen und dadurch kontrollieren (Stratifizierung). Allerdings ist dies naturgemäß nur bei beobachteten Merkmalen möglich. Beim randomisierten Zugang hingegen kann man erwarten, dass dies (im Mittel) auch für die nicht beobachteten Hintergrundvariablen gilt. Verwandt mit dem Stratifizieren ist als Alternative das (paarweise) Matching, bei dem jedem Fall aus der Treatmentgruppe ein Fall mit derselben Merkmalskombination aus der Kontrollgruppe („statistische Zwillinge“) zur Seite gestellt wird. Das Problem hierbei ist, dass schon bei wenigen Merkmalen sehr viele Kombinationen möglich sind. Auch in großen Datensätzen dürfte es daher schwierig sein, ein zweites Individuum mit der exakt gleichen Kombination der relevanten Drittvariablen zu finden.

Einen Ausweg aus diesem Problem ermöglicht das Konzept des sogenannten *Propensity Score*. Unter dem *Propensity Score* P(X) wird die Wahrscheinlichkeit für Treatment bei Individuen mit Eigenschaften X verstanden. Das zentrale Theorem von Rosenbaum & Rubin (1983) besagt, dass wenn die Bedingung der *Ignorability* konditional auf P(X) erfüllt ist (d.h. „unter Kontrolle von P“), dann dies auch konditional auf X gilt. Der Vorteil dabei ist also, dass man nicht alle möglichen X heranziehen muss, um die beiden Gruppen auszubalancieren; es genügt hierfür der Propensity Score! Auf Basis von P wird die kontrafaktische Kontrollgruppe gebildet (insbesondere durch Matching nach dem Propensity score)[3]. Eine einfache Möglichkeit, den Propensity Score empirisch zu bestimmen, sind logit/probit-Modelle. Dies ist allerdings nur mit beobachteten Variablen möglich.

[3] Als Einschränkung ist indes zu beachten, dass diese Schlussfolgerung nur für den *wahren* Propensity Score gilt, nicht notwendigerweise den faktisch vorhandenen empirischen Schätzwert.

Abbildung 4: Kausalmodell für eine „Kontrolle“ von Drittvariablen mittels Propensity Score



Konstruiert wird mit P also eine kontrafaktische Kontrollgruppe. Daraus ergibt sich ein bekanntes Paradox: Wenn das Modell zur Bestimmung des Propensity Scores „zu gut“ arbeitet, wenn also für alle Treatment-Fälle P = 1 und für alle Kontroll-Fälle P = 0 gilt, dann ist ein Matching nicht mehr möglich, da es dann keinen gemeinsamen Stützbereich mehr gibt.

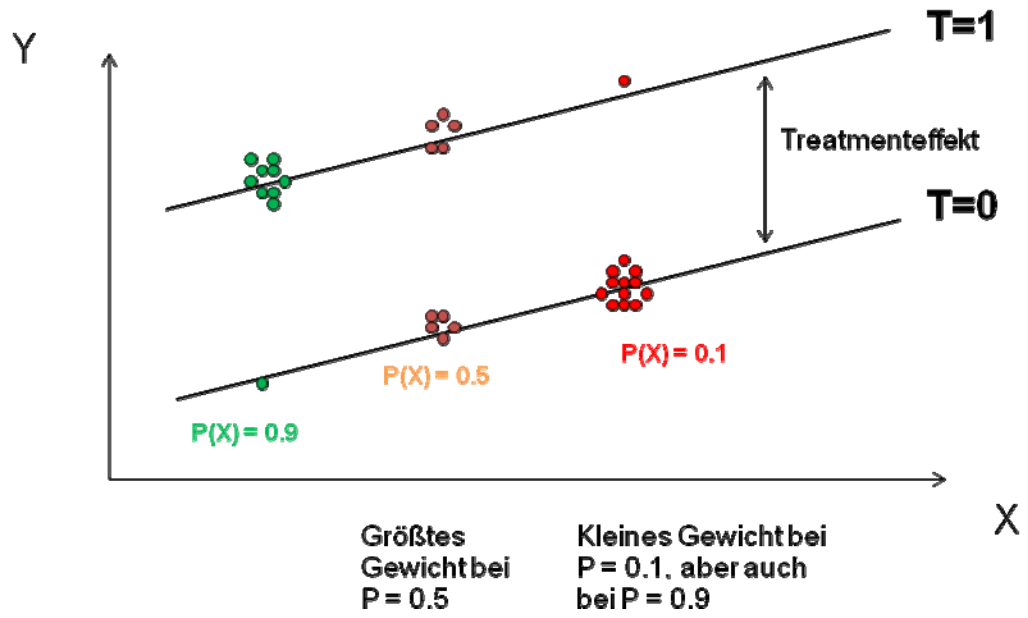
2.4 Stützbereiche und Gewichtung

Beim Vergleich der Ergebnisse ist zu beachten, dass die Verfahren auf durchaus unterschiedlichen Prinzipien basieren. Die (unbalancierte) Regression ist gekennzeichnet durch die Annahme eines funktionalen (insbesondere linearen) Zusammenhangs. Drittvariablen treten als „Kontrollvariablen im Modell der abhängigen Variablen“ in Erscheinung. Die Schätzergebnisse liefern Aussagen über den gesamten Wertebereich der Drittvariablen, etwaige Lücken im Stützbereich des Gruppenvergleichs werden modellgemäß überbrückt. Faktisch existiert häufig aber kein ausreichender gemeinsamer Stützbereich „vergleichbarer Fälle“ zwischen den beiden Gruppen. Effektschätzungen mit balancierten Daten arbeiten an sich nichtparametrisch; es müssen also keine Annahmen über ganz bestimmte funktionale Zusammenhänge gemacht werden. Allerdings erfolgt hier die Schätzung des Propensity Scores mittels eines logit-Modells. Drittvariablen treten bei der „Kontrolle der Gruppenzuordnung“ in Erscheinung. Der gemeinsame Stützbereich der beiden Gruppen wird explizit gemacht; allerdings ergibt sich bei diesem Vorgehen oft eine starke Reduktion der Fallzahlen.

Schließlich folgt aus den Verfahren eine unterschiedliche Gewichtung bestimmter Arten von Fällen. Bei der (unbalancierten) Regression (Abbildung 5) werden die Fälle so verwendet, wie sie sich in der Stichprobe verteilen. Dies bedeutet naturgemäß, dass die Treatmentgruppe überdurchschnittlich aus Fällen mit „hohem Treatmentrisiko“ und die Kontrollgruppe

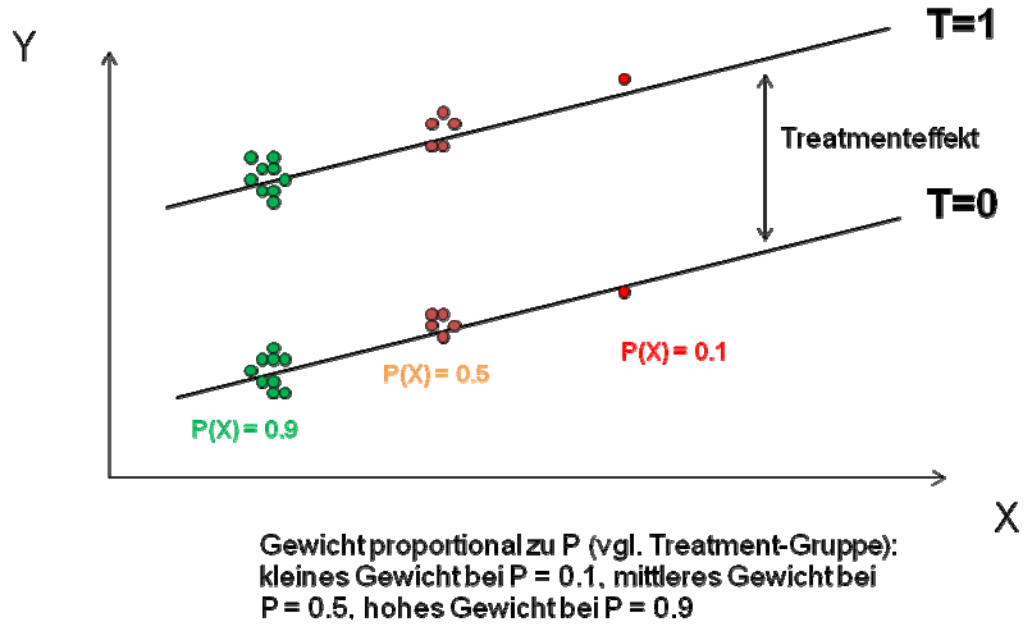
überdurchschnittlich aus Fällen mit „niedrigem Treatmentrisiko“ besteht. Entscheidend für die Bestimmung des Treatmenteffekts ist aber die Differenz zwischen den Fällen aus beiden Gruppen. Sowohl bei den Fällen mit hohen als auch mit niedrigen Risiken stehen jeweils nur sehr wenige Vergleichsfälle aus der anderen Gruppe zur Verfügung. Beim Vergleich der beiden Gruppen kommt hohes Gewicht also Merkmalskombinationen X zu, welche mit „mittleren Risiken“ eines Treatments verbunden sind.

Abbildung 5: Gewichtung von Fällen im (unbalancierten) Regressionsmodell



Anders im balancierten bzw. gematchten Datensatz (vgl. Abbildung 6). Hier werden Treatment- und Kontrollgruppe gezielt parallel besetzt, und das Propensity-Score-Matching unterscheidet zwischen dem ATT und einem möglichen Treatment-Effekt für die Kontrollgruppe. Geht man von der Treatmentgruppe aus, so besteht diese naturgemäß aus vielen Fällen, die ein „hohes Risiko“ für das Treatment haben. Da dies auch für die konstruierte Kontrollgruppe gilt, liegt also ein hohes Gewicht vor allem bei diesen Fällen.

Abbildung 6: Gewichtung von Fällen im gematchten Datensatz (ATT)



Im vorliegenden Fall repräsentiert der gemessene Statuseffekt also vor allem die Bildungsrendite, die Menschen mit hoher Studierneigung durch ihr Studium gehabt haben (im Vergleich zur Situation, dass sie nicht studiert hätten). Dabei dürfte es sich um eine in vielfacher Hinsicht selektive Teilpopulation handeln. Dies zeigt auch, dass es häufig sinnvoll ist, gruppenspezifische Treatmenteffekte zu bestimmen.

3. Eine Simulationsstudie

Diese Überlegungen sollen im Folgenden anhand von Individualdaten über individuelle Bildungskonsequenzen illustriert werden. Um sich auf die Beschreibung der reinen Verfahrensunterschiede – unabhängig von unterschiedlichen Stichproben, verfügbarer Information etc. – konzentrieren zu können, basieren die folgenden Analysen nicht auf empirischen, sondern auf simulierten Daten.

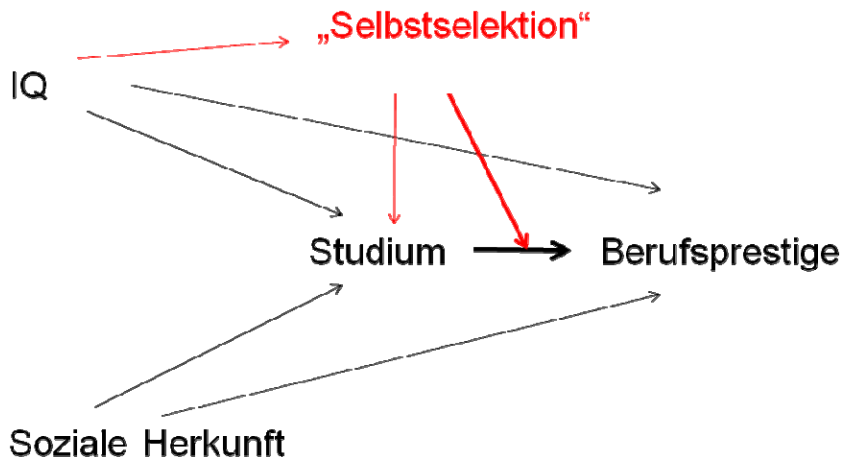
Das Simulationsmodell basiert auf den folgenden einfachen Annahmen:

- Es werden Abiturienten (mit und ohne Studium) verglichen, und es gibt auch keine weiteren formalen Einschränkungen für bestimmte Ausbildungsgänge
- Beobachtbare Determinanten des Studiums sind: Kognitive Grundfähigkeiten (IQ) und soziale Herkunft

- Beobachtbare Determinanten des beruflichen Status sind: Bildungsabschluss (Studium), IQ und soziale Herkunft
- Außerdem gibt es eine nicht beobachtbare Variable, die man als „Selbstselektion“ bezeichnen könnte. Gemeint ist damit, dass die Wahrscheinlichkeit der Absolvierung eines Studiums von den (zu erwartenden) Effekten des Studiums auf die Berufstätigkeit abhängt.

Abbildung 7 stellt das zugrundeliegende Kausalmodell dar.

Abbildung 7: Kausalmodell der Simulationsstudie



Die für die Generierung der Daten angenommenen Verteilungen und Zusammenhänge sind fiktiv, sie bewegen sich aber – soweit beobachtbar – in einer realistischen Größenordnung.

- IQ und Selbstselektion sind standardnormalverteilt und korrelieren mit r=0.7
- Als Dummyvariablen werden drei große Herkunftsschichten (OMS, MMS, UMS) mit selektiver Bildungsbeteiligung und Auswirkungen auf die berufliche Platzierung unterschieden
- außerdem gibt es jeweils standardnormalverteilte Zufallskomponenten (ZV)

Die Wahrscheinlichkeit eines Studiums für einen konkreten Fall errechnet sich als:
F (1+ IQ – UMS + OMS + Selbstselektion + ZV) mit F als dem Quantil der Verteilungsfunktion der Standardnormalverteilung; bei Werten > 0.5 gilt ein Studium als aufgenommen.

Beruflicher Status (Berufsprestige) errechnet sich als:
30+ 5*IQ- 5*UMS + 5*OMS + 10*Studium + 5*Studium*Selbstselektion[4] + 5*ZV2;

Das Modell arbeitet mit N= 1.000.000 Fällen.
Mit der entsprechenden Generierung der Daten sind für jede Person „wahre" Werte bekannt. Je nach Ausprägung der Selbstselektionsvariable schwankt die individuelle Rendite des Studiums zwischen 0 und 10. Für den Vergleich der Verfahren wird für die betreffenden Teilpopulationen jeweils die Differenz zwischen dem auf ihrer Basis gelieferten (mittleren) Schätzwert und dem (mittleren) „wahren Wert" des Bildungseffekts bestimmt. Regression und Propensity Score Matching werden jeweils mit den *gleichen* Variablen verglichen: Bei der *Regression* wird der Partialeffekt des Studium auf Berufsstatus unter Kontrolle der anderen (beobachteten) Variablen betrachtet; beim *Matching* die Gruppendifferenz des mittleren Berufsstatus, wobei das Matching (Kernelmatching, mit Normalverteilung) mit den gleichen Variablen durchgeführt wird, die in die Regression als Kontrollvariablen eingehen. Die Regressionsanalysen werden mit der Standardprozedur in STATA, die Schätzungen mittels Propensity-Score-Matching mit Hilfe des STATA-Moduls PSMATCH2, Version 3.0.0 (Leuven/Sianesi 2003) durchgeführt.

Schließlich werden die Modellparameter variiert, um mögliches unterschiedliches Verhalten der Verfahren bei unterschiedlichen Randbedingungen studieren zu können. Folgende Parameter werden variiert: (1) Der Einfluss der Selbstselektion auf die Bildungsrendite (0...max. Verdopplung). Dies bedeutet letztlich eine Veränderung der Differenz zwischen den Werten von ATT und ATU. (2) Die mittlere Wahrscheinlichkeit des „Treatments" (also letztlich die Studierendenquote). Dies hat Konsequenzen für die Gewichtung der Fälle bzw. die Größenrelation zwischen Treatment-/ Kontrollgruppe.

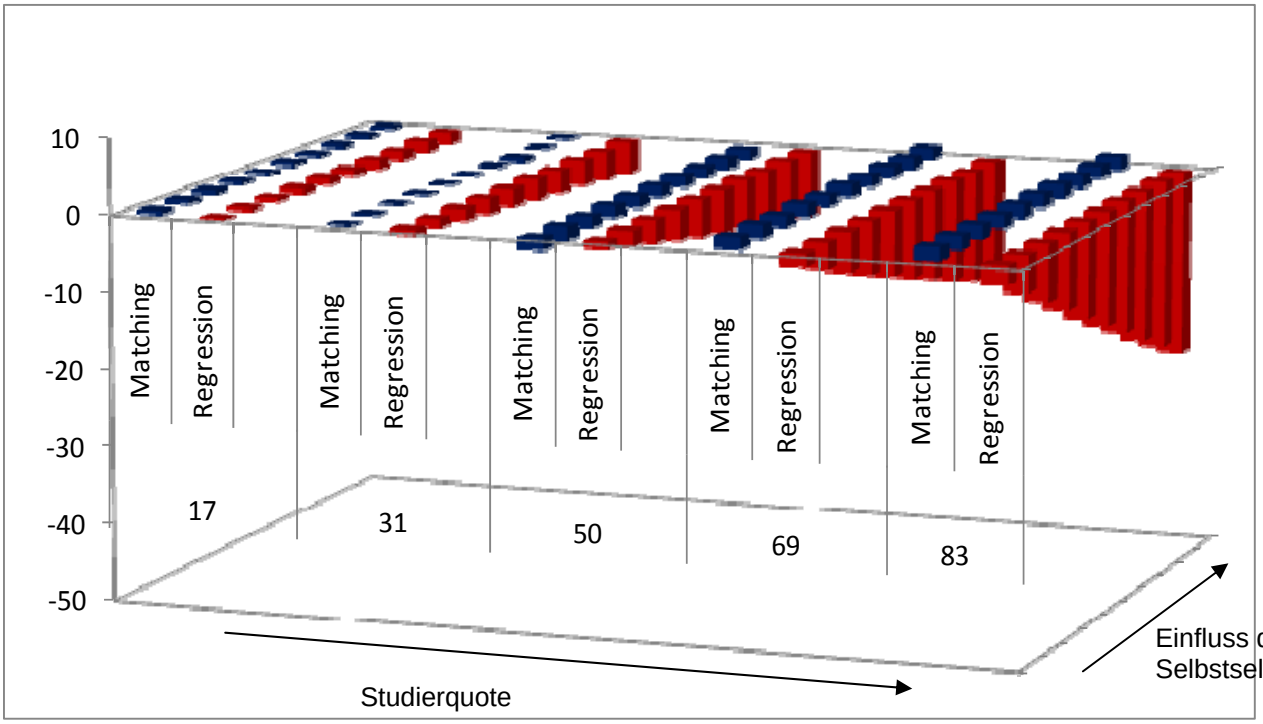
4. Ergebnisse

Zunächst wird die Güte der Schätzwerte für den ATT zwischen den beiden Verfahren unter unterschiedlichen Randbedingungen verglichen (vgl. Abbildung 8). Mit dem Matchingverfahren (das explizit zwischen ATT und ATU unterscheidet) werden die „wahren Werte" durchgängig relativ gut abgebildet. Mit dem Regressionsverfahren, das nur einen einheitlichen Schätzwert für die Gesamtpopulation ausgibt, wird die Bildungsrendite

[4] Beschränkt auf den Wertebereich +/- 2

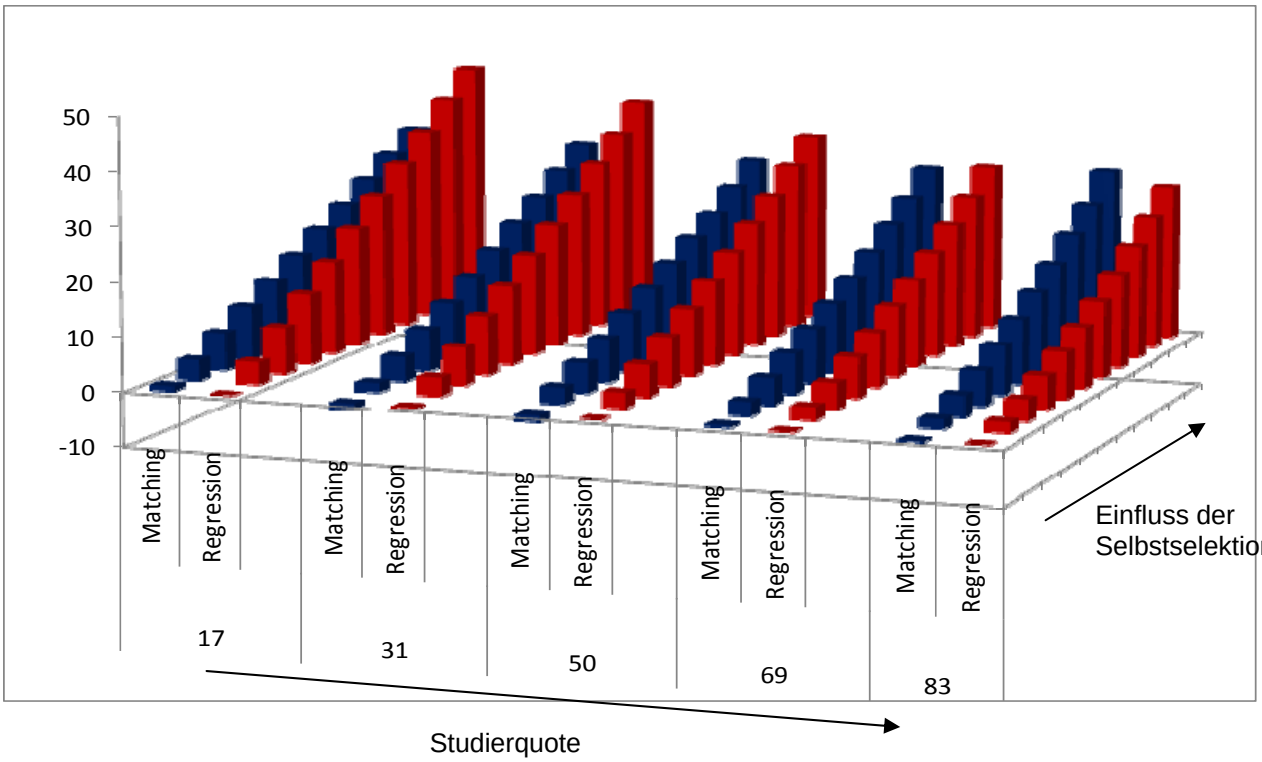
tendenziell unterschätzt, insbesondere wenn eine hohe Studierquote und ein hoher Grad von Selbstselektion im oben definierten Sinn vorliegen.

Abbildung 8: Bildungsrendite der Studierenden (ATT): Abweichungen vom "wahren" Wert in %, nach Verfahren



Beim Vergleich der Schätzungen für den ATU (Abbildung 9) fällt auf, dass die (potenzielle) Rendite eines Studiums für die Nicht-Studierenden durchgängig überschätzt wird, insbesondere bei einem hohen Grad von Selbstselektion. Dies gilt (wenn auch in unterschiedlichem Ausmaß) für beide Verfahren und auch bei verschiedenen Studierquoten. Dieses Ergebnis hat unmittelbare Konsequenzen für inhaltliche Schlussfolgerungen, da sich das Potenzial für eine weitere Bildungsexpansion ja aus den (bisher) nicht Studierenden rekrutiert. Deren zu erwartende Rendite ist also offensichtlich geringer, als es nach den „empirischen“ Ergebnissen scheint.

Abbildung 9: Bildungsrendite der Nicht-Studierenden (ATU): Abweichungen vom "wahren" Wert in %



5. Schlussfolgerungen

Die Ergebnisse zeigen, dass verschiedene Verfahren der Kausalanalyse – von denen hier exemplarisch nur zwei verglichen wurden – unter unterschiedlichen Randbedingungen offensichtlich zu systematisch unterschiedlichen Resultaten führen. Dies ist kein Zufall, denn die Verfahren unterscheiden sich u.a. im Prinzip der „Kontrolle“ von Drittvariablen und in der Gewichtung von Fällen. Das Design des Matching-Verfahrens ist expliziter auf die theoretischen Probleme der Kausalanalyse ausgerichtet. Eine Einschränkung ergibt sich aber z.B. daraus, dass als „Treatment“ jeweils nur binäre Variablen verwendet werden können. Beide Verfahren sind anfällig für Verzerrungen aufgrund unbeobachteter Faktoren. In jedem Fall unterstreichen die Ergebnisse aber die Notwendigkeit, jeweils gegenstandsspezifisch nach guten Beobachtungssituationen und Operationalisierungen zu suchen statt auf die einfache „Kontrolle“ von unzähligen Variablen in Standardregressionsmodellen zu setzen. Im konkreten Fall legen die Ergebnisse nahe, dass – beim Vorliegen von Prozessen der Effekten der Selbstselektion – die Rendite der Studierenden (ATT) im Regressionsmodell tendenziell unterschätzt, während die Rendite der Nicht Studierenden (ATU) allein durch das verwendete Analyseverfahren vermutlich häufig deutlich überschätzt wird.

In inhaltlicher Hinsicht handelt es sich bei diesem Modell natürlich um eine grobe Vereinfachung. Ob nun eine weitere Bildungsexpansion sinnvoll ist, lässt sich zudem ohne Rückgriff auf bestimmte normative Positionen nicht entscheiden. Auch haben wir von potenziellen Makro-Wirkungen abgesehen, also etwa Frage, welche Veränderungen von Bildungsrenditen zu erwarten sind, wenn nicht mehr eine Minderheit, sondern die große Mehrheit eines Jahrgangs ein Hochschulstudium absolviert. Die im konkreten Einzelfall (kontrafaktisch) zu erwartende Rendite bleibt darüber hinaus grundsätzlich unbekannt. Dennoch unterstreichen unsere Ergebnisse, wie wichtig es im Rahmen solcher Diskussionen ist, zu spezifizieren, welche Effekte genau jeweils von Interesse sind. So dürfte es gerade für Fragen der *Expansion* von großer Bedeutung sein, sich viele stärker als bisher auf die (potenziellen) Effekte für die aktuell *nicht* Studierenden (also den ATU) zu konzentrieren. Gerade vor dem Hintergrund der offensichtlich problematischen Schätzungen legt dies offensichtlich nahe, mehr Forschung zur Verbindung von (Selbst-)Selektion beim Bildungszugang und (wahrgenommenen) Bildungskonsequenzen zu betreiben. Dies gilt natürlich gerade für *empirische* Forschungen, wobei diese methodisch nicht festgelegt sind. Insofern wäre dabei zunächst durchaus auch an kleinere, stärker explorative Untersuchungen zu denken.

Literatur

Becker, Gary S. (1964): Human capital: a theoretical and empirical analysis. New York: Columbia Univ. Press.

Becker, Rolf/Hadjar, Andreas (Hg.) (2006): Die Bildungsexpansion: erwartete und unerwartete Folgen. Wiesbaden: Verlag für Sozialwissenschaften.

Gangl, Markus/DiPrete, Thomas A. (2004): Kausalanalyse durch Matchingverfahren. In: Diekmann, Andreas (Hg.): Methoden der empirischen Sozialforschung. Kölner Zeitschrift für Soziologie und Sozialpsychologie, Sonderheft 44. Wiesbaden: Verlag für Sozialwissenschaften, S. 396-420.

LaLonde, Robert (1986): Evaluating the econometric evaluations of training programs with experimental data. American Economic Review 76, S. 604-620.

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

Longford, Nicholas T. (2008): Studying Human Populations: An advanced course in statistics. Berlin: Springer.

Mincer, Jacob (1974): Schooling, experience and earnings. New York: Columbia Univ. Press.

Morgan, Stephen/Winship, Christopher (2007): Counterfactuals and causal inference. Cambridge: Cambridge Univ. Press.

Müller, Walter, 1998: Erwartete und unerwartete Folgen der Bildungsexpansion. In: Friedrichs, Jürgen, Lepsius, M. Rainer und Mayer, Karl Ulrich (Hg.): Die Diagnosefähigkeit der Soziologie. Kölner Zeitschrift für Soziologie und Sozialpsychologie, Sonderheft 38. Opladen: Westdeutscher Verlag, S. 81-112.

Rosenbaum, Paul R./Rubin, Donald B. (1983): The central role of the propensity score in observational studies for causal effects. Biometrika 70, S. 41-55.

Rubin, Donald B. (1974): Estimating Causal Effects of Treatments in Randomised and Non-Randomised Studies. Journal of Educational Psychology 66, 688-701.