

Mikroökonometrische Evaluationsstudien: Anmerkungen zu Theorie und Praxis

Michael Lechner, Fakultät für Volkswirtschaftslehre der Universität Mannheim[#]

Beitrag für Friedhelm Pfeiffer und Winfried Pohlmeier (Hrsg.): *Qualifikation, Weiterbildung und Arbeitsmarkterfolg*, ZEW-Wirtschaftsanalysen, Nomos Verlag, 1998.

Diese Version: Dezember 1997

Zusammenfassung

In diesem Aufsatz werden Methoden zur Identifikation kausaler Effekte von staatlichen Eingriffen am Beispiel von Fortbildungs- und Umschulungsprogrammen vorgestellt und deren Implikationen und Voraussetzungen diskutiert. Desweiteren werden Eigenschaften jeweils in Frage kommender Schätzmethoden erörtert. Dadurch soll dieser Aufsatz eine Hilfestellung bei der Auswahl einer für das jeweilige Problem geeignete Evaluationsmethode geben.

Abstract

The paper surveys several identification strategies for the causal effect of government intervention. The example used is public sector sponsored training. Implications of the various strategies as well as properties of some estimation procedures will be discussed.

1. Einleitung

Im Durchschnitt der Jahre 1990 bis 1996 wurden in der Bundesrepublik Deutschland jährlich etwas über 26 Mrd DM von der Bundesanstalt für Arbeit (BA) für die Programme der aktiven Arbeitsmarktpolitik ausgegeben. Dabei entfielen knapp 14 Mrd DM auf den Bereich Fortbildung und Umschulung (FuU), der in diesem Aufsatz im Mittelpunkt der Betrachtung steht.¹ Bei diesen erheblichen Ausgaben ist es erstaunlich, wie wenig über die Wirkungen der Maßnahmen sowohl aus der Sicht der einzelnen Teilnehmer als auch aus gesamtwirt-

[#] Ich danke der Deutschen Forschungsgemeinschaft für die finanzielle Unterstützung (Projekt Le 928/2). Die ausführlichen Kommentare von Martin Eichler, Bernd Fitzenberger und Marc Wellner und den Editoren haben den Aufsatz erheblich verbessert. Hinweise von Barbara Knoth haben die Darstellung deutlich verbessert. Alle verbleibenden Fehler und Unzulänglichkeiten gehen natürlich voll zu Lasten des Autors. Von meiner homepage (http://www.vwl.uni-mannheim.de/inst_vwl/lechner) können weitere Arbeitspapiere zu diesem Thema heruntergeladen werden.

¹ Quelle: Verschiedene Ausgaben der *Amtlichen Nachrichten der Bundesanstalt für Arbeit* (eigene Berechnungen). Die Gesamtzahlen enthalten Ausgaben für FuU (einschließlich Einarbeitungszuschüsse), Arbeitsbeschaffungsmaßnahmen, Kurzarbeit, Maßnahmen nach §249h und Vorruhestand.

schaftlicher Sicht bekannt ist.² Dies steht im krassen Gegensatz zu der Situation in den USA. Obwohl die dortigen Arbeitsmarktprogramme von wesentlich geringerem relativem Umfang als die Programme der BA sind, wurden doch erhebliche Ressourcen darauf verwendet, deren Wirkung auf die individuelle Situation der jeweiligen Teilnehmer zu verstehen. Vom Standpunkt des Autors ist hierbei besonders erfreulich, daß die Ergebnisse wiederum in die Politik zurückgefließen sind und zu einer Veränderung sowohl des Volumens und als auch der Ausgestaltung der Programme geführt haben (siehe z.B. Bloom et al., 1997, Heckman et al., 1997, und Lalonde, 1995).

Ein Nebenprodukt dieser erheblichen Anstrengungen ist eine bedeutende wissenschaftliche, empirisch-ökonometrische Literatur, in der sowohl vom methodischen Standpunkt als aus der Perspektive der praktischen Umsetzung diskutiert wird, wie ökonometrische Evaluationen am besten durchgeführt werden können. Beispielhaft seien hier die Arbeiten von Ashenfelter (1978), Heckman, Robb (1985), Card, Sullivan (1988), Heckman, Hotz (1989) und Angrist et al. (1996) genannt. In jüngster Zeit sind auch für die Bundesrepublik Deutschland (BRD) mikroökonometrische Arbeiten erschienen, die auf dieser Literatur aufbauen bzw. diese erweitern.³ Im großen und ganzen muß man allerdings feststellen, daß unser Wissen über die Wirksamkeit beispielsweise von FuU bis dato nur als sehr gering eingestuft werden kann. Diese Feststellung trifft insbesondere dann zu, wenn man Fragen nach den Effekten der FuU für unterschiedliche Ausgestaltungen der Programme oder für unterschiedliche Teilnehmergruppen beantworten möchte. In der BRD dürfte es daher noch eine geraume Zeit dauern, bis die Ergebnisse dieser Studien in die Politik einfließen und zu Änderungen künftiger Programme der aktiven Arbeitsmarktpolitik führen werden.

Dieser Aufsatz hat sich zum Ziel gesetzt, wichtige Methoden zur Identifikation von kausalen Effekten von staatlichen Eingriffen anhand des Beispiel von Fortbildungs- und Umschulungsprogrammen (*Training*) vorzustellen und deren Implikationen sowie auch teilweise die Eigenschaften jeweils in Frage kommender Schätzmethoden zu diskutieren.⁴ Zu diesem Zweck wird im folgenden Abschnitt zunächst ein Überblick über die grundsätzliche Problematik mikroökonometrischer Evaluierungsstudien gegeben. Unter „mikroökonometrischen Evaluationsstudien“ sollen hier Analysen verstanden werden, die die Wirkungen eines bestimmten Programms aus der Sicht der Programmteilnehmer und Nichtteilnehmer empirisch analysieren. Sie versuchen, die Effekte eines Programms durch einen (hypothetischen) Vergleich mit einem „Zustand der Welt“ ohne dieses Programm abzuschätzen. Der Schwerpunkt liegt hier bei sogenannten nichtexperimentellen

² Dies betrifft natürlich nicht die von der Bundesanstalt für Arbeit (BA) verwendeten sogenannten Entlastungseffekte, definiert als die Anzahl der Personen, die durch die Teilnahme an der Maßnahme dem Arbeitsmarkt temporär nicht zur Verfügung stehen.

³ Siehe z.B. Hübler (1997) und die Übersichtsartikel von Hujer et al. (1996) und von Fitzenberger und Prey in diesem Band.

⁴ In gewissem Sinne kann dieser Aufsatz als eine Ergänzung der sehr lesenswerten und weitergehenden Arbeit von Schnellhaaß und Schubert (1992) verstanden werden.

Evaluationen, da in diesem Bereich in der BRD im Gegensatz zu den USA keine sozialen Experimente durchgeführt wurden. Es wird diskutiert, was eigentlich die typischerweise interessierenden Fragen einer Evaluationsstudie sind (Beispiel: Wie groß ist der mittlere Effekt für die Teilnehmer durch die Programmteilnahme?). Auch wird eine Notation eingeführt, die es erlaubt, einen bestimmten Effekt eines Programmes präzise zu definieren. Es folgen Überlegungen zu der Frage, anhand welcher Indikatoren die Messung des Erfolgs der Programme erfolgen kann (Beispiel: Einkommen). Abschließend werden die Annahmen diskutiert, die für eine sinnvolle Anwendung dieser Notation notwendig sind. Der dritte Abschnitt beschäftigt sich unter dem Stichwort Identifikation mit den Bedingungen, die notwendig sind, um aus einem vorhandenen Datensatz die zuvor definierten Effekte der Programme schätzen zu können und diskutiert die Eigenschaften einiger wichtiger, aber noch nicht weit verbreiteter Schätzmethoden. Abschnitt 4 enthält die Schlußfolgerungen und einen Ausblick für die zukünftige Forschung.

2 Evaluation und Kausalität: Ein Modellrahmen

2.1 Notation und interessierende Größen

Um Aussagen über die Wirkung einer Trainingsmaßnahme auf ein einzelnes Individuum treffen zu können, ist es notwendig, formal korrekt zu beschreiben, was man unter einer (kausalen) Wirkung versteht. Insbesondere gilt es, den kausalen Effekt einer Maßnahmenteilnahme von einer Korrelation der Maßnahmenteilnahme mit einer Ergebnisvariable (z.B. Arbeitsmarktstatus) zu unterscheiden.⁵ Für ökonometrische Evaluationen hat sich weitgehend das Modell potentieller Ergebnisse (Roy, 1951, Rubin, 1974) durchgesetzt. Die Grundidee des Roy-Rubin-Modells besteht in einem Vergleich des Ergebnis eines Individuums bei einer Maßnahmenteilnahme (Y^t) mit dem Ergebnis bei Nichtteilnahme (Y^n). Als kausaler Effekt kann dann zum Beispiel die Differenz der potentiellen Ergebnisse ($Y^t - Y^n$) aufgefaßt werden. Entscheidend ist also der Vergleich zweier Zustände, die die „Welt“ vollständig beschreiben, von denen aber immer nur ein Zustand beobachtbar und der andere dann per definitionem unbeobachtbar sein muß.⁶ Damit man hier von Kausalität sprechen kann, muß die Annahme getroffen werden, daß die individuellen potentiellen Ergebnisse nicht von dem Teilnahmeverhalten anderer Individuen abhängen (*stable unit treatment value*-Annahme, SUTVA).⁷

Es ist offensichtlich, daß es aufgrund der Unbeobachtbarkeit eines der beiden Ereignisse unmöglich ist, einen individuellen kausalen Effekt abzuschätzen. Hat man jedoch eine größere

⁵ Siehe z.B. Holland (1986) und Sobel (1994) für ausführliche Diskussionen der Kausalitätskonzepte in Ökonometrie, Statistik und anderen Wissenschaften.

⁶ Eine Erweiterung auf mehrere Zustände ist möglich, erhöht aber den Aufwand an notwendiger Notation stark und soll daher hier unterbleiben, da die meisten wichtigen Fragen anhand des einfachen Modells mit zwei Zuständen diskutiert werden können.

⁷ Ist SUTVA in dieser strengen Form nicht erfüllt, so erweitert sich die Anzahl der potentiellen Ergebnisse stark und eine Identifikation eines kausalen Effekts ist erheblich erschwert (siehe Abschnitt 2.3).

Stichprobe mit Teilnehmern und Nichtteilnehmern an der Maßnahme zur Verfügung, so ist es in bestimmten Fällen, die in den folgenden Abschnitten näher diskutiert werden sollen, möglich, mittlere kausale Effekte für Untergruppen möglicher und tatsächlicher Teilnehmer zu identifizieren. Die Merkmale derartiger Untergruppen sollen allgemein mit der Variable Z beschrieben werden. Entsprechend wird eine bestimmte Ausprägung von Z im folgenden mit z dargestellt. Bezeichnet die Variable S das Teilnahmeverhalten eines Individuums (1 für Zustand t , 0 für Zustand n), so können drei unterschiedliche kausale Effekte definiert werden:⁸

$$\gamma_0(z) := E(Y^t - Y^n \mid Z = z) = E(Y^t \mid Z = z) - E(Y^n \mid Z = z), \quad (1)$$

$$\theta_0(z) := E(Y^t - Y^n \mid Z = z, S = 1) = E(Y^t \mid Z = z, S = 1) - E(Y^n \mid Z = z, S = 1), \quad (2)$$

$$\xi_0(z) := E(Y^t - Y^n \mid Z = z, S = 0) = E(Y^t \mid Z = z, S = 0) - E(Y^n \mid Z = z, S = 0). \quad (3)$$

$\gamma_0(z)$ steht für den erwarteten Maßnahmeneffekt eines zufällig aus einer Gruppe mit Charakteristika z ausgewählten Individuums. Entsprechend stehen $\theta_0(z)$ und $\xi_0(z)$ jeweils für den Effekt eines aus der Gruppe der Teilnehmer bzw. aus der Gruppe der Nichtteilnehmer ausgewählten Individuums. In vielen Evaluationsstudien steht der Parameter $\theta_0(z)$ im Vordergrund. $\theta_0(z)$ gibt den Maßnahmenerfolg für Personen an, die tatsächlich am Programm teilgenommen haben und kann somit als ein Erfolgsindikator des Programm aufgefaßt werden. Oft wäre es wünschenswert, mehr über die Verteilung der individuellen kausalen Effekte zu erfahren als nur den Erwartungswert. Die meisten anderen statistischen Maße für die Verteilung der kausalen Effekte (Varianz, Median, etc.) haben jedoch den aus ökonometrischer Sicht gravierenden Nachteil, daß sie von Eigenschaften der gemeinsamen Verteilung von (Y^t, Y^n) abhängen und nicht allein anhand der marginalen Verteilungen von Y^t und Y^n identifizierbar sind, wie dies beim Erwartungswert der Fall ist. Da jedoch für jedes Individuum nur Y^t oder Y^n beobachtet werden kann, ist die Identifikation von derartigen Parametern der gemeinsamen Verteilung unmöglich.⁹

2.2 Variablen zur Messung der Ergebnisse der Maßnahmen?

In jeder empirischen Arbeit stellt sich natürlich die Frage, mit welchen Variablen die Ergebnisse Y gemessen werden können und sollen. Für diese Betrachtungen ist es nützlich,

⁸ $E(Y^t - Y^n \mid Z = z, S = s)$ sei der Erwartungswert in einer Gruppe der Bevölkerung, die sich durch die Charakteristika definiert, die nach dem „|“ angegeben sind. Damit die Effekte sinnvoll definiert sind, muß Z 'exogen' sein. Exogenität ist hier in dem Sinne definiert, daß die beobachtbare Realisation von Z nicht davon abhängt, welcher der beiden Zustände von S realisiert wird, d. h. es gilt: $Z = Z^t = Z^n$. Diese Definition entspricht nicht unbedingt anderen in der Ökonometrie verwendeten Exogenitätsbegriffen (vgl. auch Holland, 1986).

⁹ Inwieweit einige dieser Maße im Falle eines sozialen Experiments zumindest auf ein Intervall eingeschränkt werden können, zeigen Clements et al. (1994). Die Intervallbreite ist allerdings typischerweise zu groß, um auch nur bezüglich des Vorzeichens eindeutige Ergebnisse zu erhalten.

eine explizite Zeitdimension einzuführen. Hier soll aus Vereinfachungsgründen davon ausgegangen werden, daß die Maßnahme nur zu einem einzigen bestimmten Zeitpunkt begonnen werden kann und an einem einzigen späteren Zeitpunkt beendet wird.¹⁰ Damit ist es ohne Einschränkung der Allgemeinheit, möglich, die zeitliche Messung relativ zum Datum des Beginns der Maßnahmen durchzuführen. Die potentiellen Ergebnisse können dann als Vektoren $Y^t = \{Y_{-T}^t, \dots, Y_{-2}^t, Y_{-1}^t, Y_0^t, Y_1^t, Y_2^t, \dots, Y_T^t\}$ und $Y^n = \{Y_{-T}^n, \dots, Y_{-2}^n, Y_{-1}^n, Y_0^n, Y_1^n, Y_2^n, \dots, Y_T^n\}$ geschrieben werden. Oft, aber nicht immer, wird man davon ausgehen können, daß die Werte der Zielvariablen, die zeitlich vor dem Programm angesiedelt sind, nicht von der späteren Programmteilnahme beeinflusst werden. Somit sind sie als Ergebnisvariablen nicht relevant.¹¹ Ob man für die Evaluation den Zeitraum der Maßnahme mitberücksichtigt (Y_0) oder erst die Ergebnisse nach Maßnahmenende analysiert ($Y_1, Y_2, \dots$), hängt davon ab, ob man die direkten Effekte während der Maßnahme für wirtschaftspolitisch relevant hält. Wird beispielsweise eine Kosten-Nutzen-Analyse angestrebt, so sind sicher alle Ergebnisse relevant. Geht man aber der einfacheren Frage nach, ob es die Teilnehmer durch die Maßnahme am Arbeitsmarkt leichter haben, so ist nur die Zeit nach der Maßnahme von Bedeutung. Hat man diese Frage geklärt, so kann man den Maßnahmeerfolg für jeden Zeitpunkt einzeln betrachten und somit auch ein Bild über die Dynamik der Wirkungen gewinnen. Alternativ kann man auch die Wirkung über mehrere Zeitpunkte zusammenfassen, z.B. in einen mittleren bzw. kumulierten Effekt über die Zeit. Dieses Vorgehen hat den insbesondere bei kleiner Stichprobengröße wichtigen Vorteil, daß durch die zusätzliche Mittelwertbildung die Varianz der geschätzten Effekte reduziert wird.

Typischerweise wird Y mit Variablen quantifiziert, die den Nutzen des Programms für das einzelne Individuum beschreiben können. In fast allen Arbeiten spielt daher das Periodeneinkommen eine wichtige Rolle (siehe z.B. Ashenfelter, 1978), aber auch das Arbeitslosigkeitsrisiko – gemessen entweder als Periodenarbeitslosigkeitswahrscheinlichkeit (z.B. Card, Sullivan, 1988) oder als Abgangsrate aus Arbeitslosigkeit (z.B. Gritz, 1993, Hujer et al. 1997 a, b)¹² – wird häufig berücksichtigt. Arbeitslosigkeit ist dann besonders relevant, wenn einige öffentlich geförderte Programme nicht die Erhöhung des Einkommens zum Ziel haben, sondern explizit eine Senkung des Arbeitslosigkeitsrisikos.¹³ Alternativ oder zusätzlich kann man andere Arbeitsmarktindikatoren verwenden, in denen die Arbeitszeit implizit berücksichtigt ist, wie z.B. Vollerwerbstätigkeit, Teilzeiterwerbstätigkeit (Lechner, 1996 a, b,

¹⁰ Die Auswirkungen des Fehlens dieser Vereinfachung wird am Ende von Abschnitt 3.4 angesprochen.

¹¹ Ein Effekt des Programms vor dem Zeitpunkt der tatsächlichen Teilnahme könnte zum Beispiel dann auftreten, wenn eine Person um in ein Programm gelangen zu können, ein bestimmtes Verhalten an den Tag legt, das ihr die gewünschte Programmteilnahme erst ermöglicht (z.B. eine Arbeitslosigkeitsmeldung), oder sie ihr Verhalten aufgrund der Antizipation einer Teilnahme verändert.

¹² Siehe Ham und Lalonde (1996) für die besonderen ökonometrischen Probleme, die mit der Verwendung von Abgangsraten bzw. Intervall-Längen bestimmter Arbeitsmarktzustände als Erfolgsindikator verbunden sind.

¹³ Die Bundesanstalt für Arbeit verwendet zum Beispiel den Anteil der Empfänger von Arbeitslosenunterstützung oder Arbeitslosenhilfe im Quartal, bzw. 6 Monate nach Abschluß einer FuU als Erfolgskriterium (Vgl. für FuU das Arbeitsförderungsgesetz, AFG, § 36 (P.3) in Verb. mit § 2 (P.1, 4, 5, 6) und § 42 (P. 1, 2a, 2b)).

c), Stundenlöhne (Fitzenberger, Prey, 1996), oder die Arbeitszeit selbst. In den wenigsten Studien wird allerdings berücksichtigt, daß diese Arbeitsmarktzustände über die unterschiedliche Verfügbarkeit von Freizeit einen Effekt auf den Nutzen haben (Greenberg, 1997). Insbesondere Einkommensangaben, aber auch die Arbeitszeit haben den Nachteil, daß sie oft mit erheblicher Varianz gemessen werden. Daher sind auch andere Maße für die Qualität des Arbeitsplatzes, wie z.B. die Stellung im Beruf oder ähnliches, von Interesse (Lechner, 1995 a). In den meisten Evaluierungsstudien steht nur ein relativ kurzer Zeitraum nach Beendigung der Maßnahme zur Evaluation zur Verfügung, so daß nur kurz- bis mittelfristige Effekte identifizierbar sind. Hier können die Einschätzungen der Individuen über ihre Arbeitsmarktsituation in der Zukunft helfen, zumindest individuell erwartete Effekte über die Stichprobenperiode hinaus abzuschätzen.¹⁴

2.3 Probleme des Modellrahmens

Der oben vorgestellte Ansatz wird in allen dem Autor bekannten, auf Individualdaten basierenden Evaluationsstudien explizit oder implizit verwendet. Daher erscheint es sinnvoll, vor der Diskussion der Identifikation und der Schätzung der Effekte innerhalb dieses Modellrahmens (Abschnitt 3), auf die Probleme des Modellrahmens selbst kurz einzugehen.

Die Annahme, daß die potentiellen Ergebnisse eines Individuums nicht von der Teilnahmeentscheidung anderer Individuen beeinflusst werden (SUTVA), ist entscheidend für die kausale Interpretation. Diese Annahme ist immer dann zweifelhaft, wenn z.B. über den Marktmechanismus Verbindungen zwischen den einzelnen Marktteilnehmern bestehen. Beispielsweise kann man sich gut vorstellen, daß Y^i , z.B. der Stundenlohn, für eine Person, die in einen anderen Beruf umgeschult wird, davon abhängt, wieviele Personen insgesamt in diesen Beruf umgeschult werden.¹⁵ Es wird deutlich, daß SUTVA insbesondere immer dann zum Problem werden kann, wenn Arbeitsmarktprogramme die Angebots-Nachfrage-Relationen spürbar verändern. Erscheint dies bei relativ kleinen Programmen weitgehend unproblematisch, so kann dies bei groß angelegten Programmen, wie z.B. in den neuen Bundesländern nach der Vereinigung, zum Problem werden.¹⁶ Jedoch ist dem Autor dieses Beitrags kein Ansatz bekannt, der in solch einem Fall mit plausiblen alternativen Annahmen oder einer modifizierten SUTVA versucht hat, die Effekte großer Programme zu quantifizieren.¹⁷ In vielen Fällen erscheint allerdings selbst bei einer offensichtlichen Verletzung von SUTVA die mikroökonomische Analyse immer noch sinnvoll, da sie unter

¹⁴ Im sozio-ökonomischen Panel (GSOEP) wird zum Beispiel nach dem erwarteten Arbeitsmarktstatus in den nächsten zwei Jahren gefragt. Eine Komplikation bei Verwendung dieser Variablen besteht allerdings darin, daß die Art der Fragestellung oft vom Arbeitsmarktstatus zum Zeitpunkt der Befragung abhängig ist.

¹⁵ Daher wird SUTVA in der ökonomischen Evaluationsliteratur auch als Abwesenheit von 'Allgemeinen-Gleichgewichts-Effekten' (*no general equilibrium effects*) bezeichnet. Der Begriff der Abwesenheit von makroökonomischen Effekten ist ebenfalls synonym (siehe Garfunkel, Manski, und Michalopoulos, 1992).

¹⁶ Einen Überblick über die Programme in den neuen Bundesländern geben z.B. Eichler und Lechner (1996).

¹⁷ Es hat sich leider eher das Vorgehen durchgesetzt, SUTVA zu ignorieren und die Ergebnisse so zu interpretieren, als ob SUTVA gültig wäre.

Abschätzung der Wirkung der allgemeinen Gleichgewichtseffekte noch zumindest Schranken für die wahren Effekte liefern sollte.

3 Identifikation und Schätzung

3.1 Vorbemerkungen

Dieser Abschnitt wird sich auf die Identifikation des Effektes der Maßnahmen auf die Maßnahmenteilnehmer beschränken. Für Teilnehmer mit Charakteristika z wurde dieser Effekt in Gleichung (2) mit $\theta_0(z)$ bezeichnet. Entsprechend läßt sich der erwartete Effekt für eine Person, die aus der Gruppe der Teilnehmer zufällig ausgewählt wurde, definieren:

$$\theta_0 = E_Z[\theta_0(Z) | S = 1] := E_Z[E(Y^t - Y^n | Z, S = 1) | S = 1]. \quad (2')$$

Da es konzeptionell keine Unterschiede bei der Schätzung und Identifikation von θ_0 und $\theta_0(z)$ gibt, werden im folgenden θ_0 und $\theta_0(z)$ äquivalent verwendet.¹⁸

Das Identifikationsproblem wird deutlich, wenn man die Funktionen $g^t(z) := E(Y^t | Z = z, S = 1)$ und $g^n(z) := E(Y^n | Z = z, S = 0)$ definiert, und θ_0 und $\theta_0(z)$ in termini dieser Funktionen schreibt:

$$\begin{aligned} \theta_0(z) &= g^t(z) - E(Y^n | Z = z, S = 1) \\ &= g^t(z) - g^n(z) + [g^n(z) - E(Y^n | Z = z, S = 1)] \\ &= g^t(z) - g^n(z) + \lambda(z). \end{aligned} \quad (4)$$

$\lambda(z)$ bezeichnet dabei den sogenannten Selektionseffekt. Die Funktionen $g^t(z)$ und $g^n(z)$ sind identifiziert und können mit einfachen Methoden geschätzt werden. Es kann z.B. der Mittelwert der Y^t für die Individuen mit $S=1$, oder der Mittelwert der Y^n für die Nichtteilnehmer berechnet werden. Alternativ können die bedingten Erwartungswerte auch mit Hilfe von linearen, nichtlinearen oder nichtparametrischen Regressionsmethoden jeweils in der Stichprobe der Teilnehmer und der Nichtteilnehmer geschätzt werden. $E(Y^n | Z = z, S = 1)$ und somit auch $\lambda(z)$ können ohne weitere Annahmen nicht identifiziert werden, da von keinem Teilnehmer Y^n beobachtet werden kann. Y^n wird in diesem Fall auch als *counter-factual outcome* bezeichnet. Daher versucht ein Großteil der methodischen

¹⁸ Sei $\hat{\theta}_0(z)$ eine konsistente Schätzung für $\theta_0(z)$ so erhält man mit schwachen Regularitätsannahmen mit $\hat{\theta}_N = (1 / \sum_{i=1}^N s_i) \sum_{i=1}^N \hat{\theta}_N(z_i) s_i$ einen konsistenten Schätzer für θ_0 .

Evaluationsliteratur, plausible Annahmen zu finden, um $E(Y^n | Z = z, S = 1)$ aus den Ereignissen ($y^t, s=1$) und vor allem aus ($y^n, s=0$) zu identifizieren.

Bevor in den folgenden Abschnitten näher auf derartige Annahmen eingegangen wird, sollen noch die Arbeiten von Manski (1990, Theorie) und Manski, Sandefur, McLanahan und Powers (1992, Anwendung) angesprochen werden. Manski's (1990) Idee basiert darauf, daß unter der Annahme, daß Y^n in einem endlichen Intervall definiert ist (z.B. 0/1 bei einer Indikatorvariablen), zwar ohne zusätzliche Annahmen keine Punktschätzung von $\theta_0(z)$ möglich ist, jedoch sehr wohl ein Intervall angegeben werden kann, in dem $\theta_0(z)$ liegen muß. Man sieht dies am einfachsten, wenn man von einer Indikatorvariable ausgeht. $\theta_0(z)$ muß auf das Intervall $[-1, +1]$ beschränkt sein, da die in Gleichung (4) eingehenden Erwartungswerte $E(Y^t | Z = z, S = 1) [=g^t(z)]$ und $E(Y^n | Z = z, S = 1)$ jeweils auf das Intervall $[0,1]$ beschränkt sind. Kennt man nun aber $g^t(z)$, bzw. hat man es konsistent geschätzt, so können die Intervallbreiten halbiert werden und $\theta_0(z)$ liegt dann zwischen $[g^t(z)-1, g^t(z)]$.¹⁹ Zwar können die Intervalle durch weitere Annahmen eingeschränkt werden, jedoch zeigt Lechner (1996 c) für den Fall von FuU in den neuen Bundesländern mit den Daten des Mikrozensus 1993, daß sich diese Intervalle nur mit unplausiblen Annahmen weit genug reduzieren lassen, um damit wirtschaftspolitisch relevante Größenordnungen für $\theta_0(z)$ zu erhalten. Eine Lektion aus dieser Literatur ist sicher, daß es ohne hinreichendes Wissen über die Art und Weise, wie Individuen in die Maßnahmen hineinselektiert werden, unmöglich ist, die Effekte hinreichend genau zu quantifizieren. Die für eine exakte ökonometrische Untersuchung notwendigen (identifizierenden) Annahmen, die sich (leider!) statistisch nicht testen lassen, können nur durch solches Wissen gerechtfertigt werden. Stehen solche Informationen und die entsprechenden Daten nicht in ausreichendem Maße zur Verfügung, so sind wirtschaftspolitische Empfehlungen höchst problematisch. Die unterschiedlichen Strategien zur Identifikation werden in den folgenden Abschnitten ausführlich diskutiert.

3.2 Der Vorher-Nachher Vergleich

Der sogenannte Vorher-Nachher Vergleich (*Before-After Comparison*, BAC) war und ist eine oft verwendete Methode zur Berechnung von θ_0 . Der BAC beruht auf der Idee, daß jedes Individuum vor der Trainingsteilnahme ein Nichtteilnehmer ist. Somit ist für jede Person ein Y^n beobachtbar. Somit gewinnt man aus dem Vergleich des Y der Person vor Maßnahmeteilnahme mit dem Y nach der Teilnahme ein Maß für die Wirkung der Maßnahme. Dieses Vorgehen hat den Vorteil, daß nur eine Stichprobe der Teilnehmer benötigt wird.

BAC kann jedoch nur unter einer starken Annahme den Effekt θ_0 identifizieren. Betrachten wir der Einfachheit halber eine Welt mit nur einem Zeitpunkt vor und nach der FuU. Somit

¹⁹ Allerdings läßt sich hiermit die Hypothese $\theta_0(z)$ sei gleich null auch mit einer noch so großen Stichprobe nicht verwerfen.

lassen sich die Ergebnisse schreiben als $\{Y_{-1}^t, Y_1^t\}$ und $\{Y_{-1}^n, Y_1^n\}$. Bezeichne θ_0^{BAC} den unter der BAC- Annahme identifizierten Effekt:

$$\begin{aligned}\theta_0^{BAC} &= E(Y_1^t | S = 1) - E(Y_{-1}^t | S = 1) \\ &= g_1^t - g_{-1}^t ,\end{aligned}\tag{5}$$

mit $g^t = E_Z[g^t(Z) | S = 1]$. Bezeichne nun B^{BAC} die Abweichung von θ_0^{BAC} vom kausalen Effekt θ_0 , so erhält man:

$$B^{BAC} = \theta_0^{BAC} - \theta_0 = E(Y_1^n | S = 1) - E(Y_{-1}^t | S = 1) .\tag{6}$$

Daher lautet die dem BAC-Schätzer zugrunde liegende Annahme wie folgt:

$$B^{BAC} = 0 \Leftrightarrow E(Y_1^n | S = 1) = E(Y_{-1}^t | S = 1) .\tag{BAC}$$

Die Plausibilität dieser Annahme hängt von mehreren Faktoren ab. Zum einen dürfen die Y_{-1}^t nicht schon durch eine eventuell antizipierte Trainingsteilnahme beeinflusst werden, so daß $E(Y_{-1}^t | S = 1) = E(Y_{-1}^n | S = 1)$ plausibel erscheint. Zum anderen wird Zeitinvarianz vorausgesetzt. Würde z.B. eine Person in $t = -1$ arbeitslos, so dürfte sie keine Suchaktivitäten zeigen, da sich dadurch Y_1^n systematisch gegenüber Y_{-1}^n verbessern würde, und die Wirkung der Maßnahme überschätzt würde ($\theta_0^{BAC} > \theta_0$). Auch sind systematische Veränderungen auf dem Arbeitsmarkt ausgeschlossen. Wie dieses Beispiel schon zeigt, ist BAC in vielen Fällen als wenig plausibel anzusehen.

3.3 Zufällige Teilnahme (soziale Experimente)

In diesem und den folgenden Abschnitten wird davon ausgegangen, daß zusätzlich zu der Gruppe der Teilnehmer auch noch eine sogenannte Kontrollgruppe, also eine Gruppe von Nichtteilnehmern, beobachtet werden kann. Eine Kontrollgruppe läßt sich in besonders einfacher Art und Weise verwenden, wenn die Aufteilung der Individuen in Kontrollgruppe und Teilnehmergruppe unabhängig vom potentiellen Nichtteilnehmerergebnis Y^n ist:

$$Y^n \amalg S .\tag{IA}$$

$\amalg$ steht für Unabhängigkeit und IA für *Independence Assumption*. Der mit IA identifizierbare Effekt (θ_0^{IA}) lautet:

$$\theta_0^{IA} = E(Y^t | S = 1) - E(Y^n | S = 0) = g^t - g^n ;\tag{7}$$

$$B^{IA} = E(Y^n | S = 1) - g^n; \quad (8)$$

$$B^{IA} = 0 \Leftrightarrow g^n = E(Y^n | S = 1). \quad (IA')$$

Gleichung (8) zeigt, daß die schwächere Annahme IA' hinreichend ist, um θ_0 zu identifizieren.²⁰ IA oder IA' ist insbesondere im Falle (idealtypischer) sozialer Experimente plausibel, da dann die Zugehörigkeiten zu den Gruppen der Teilnehmer oder Nichtteilnehmer mit Hilfe eines Zufallsgenerators bestimmt wird.²¹

Da aber für die BRD keine Daten aus einem größeren sozialen Experiment im Bereich der Arbeitsmärkte zur Verfügung stehen, ist diese Annahme in der deutschen Evaluationspraxis von geringer Bedeutung. In nichtexperimentellen Situationen erkennt man oft schon bei einem ersten Blick auf die deskriptiven Statistiken für die Teilnehmer und die Nichtteilnehmer, daß es einige Variablen gibt, die sowohl mit Y^n als auch mit S korreliert sind (z.B. der Bildungsstand). Somit ist IA nicht plausibel.

3.4 Zufällige Auswahl der Teilnehmer mit identischen Charakteristika

3.4.1 Identifikation

Eine Verallgemeinerung der Unabhängigkeitsannahme ist die Annahme der bedingten Unabhängigkeit für einen gegebenen Wert eines Vektors beobachtbarer Charakteristika X (*Conditional Independence Assumption*, CIA):

$$Y^n \perp\!\!\!\perp S | X = x \quad \forall x \in \mathcal{X}. \quad (CIA)$$

$\mathcal{X}$ bezeichne den Raum von Charakteristika, für den CIA Gültigkeit hat. CIA impliziert, daß Y^n und S für gegebene Charakteristika x unabhängig sind. Damit diese Annahme plausibel ist, muß man alle Variablen kennen und im Datensatz beobachten, die Y^n und S gemeinsam beeinflussen. Daher ist sowohl ein sehr umfangreiches Wissen über den Teilnahmeprozess und die Bestimmungsgründe von Y^n , als auch ein sehr informativer Datensatz für die empirische Analyse notwendig. Der mit CIA identifizierbare Effekt (θ_0^{CIA}) ist in Gleichung (9) definiert:

$$\theta_0^{CIA} = g^t - E[g^n(X) | S = 1]; \quad (9)$$

$$g^n(x) := E(Y^n | X = x, S = 0) \stackrel{CIA}{=} E(Y^n | X = x, S = 1) \stackrel{CIA}{=} E(Y^n | X = x);$$

²⁰ Die strengere Annahme IA hat gegenüber IA' den Vorteil, daß bei Gültigkeit von IA für Y^n alle funktionalen Formen von Y^n die Annahme IA' erfüllen. IA' wird als *mean independence* bezeichnet.

²¹ Siehe Burtless (1995) und Heckman und Smith (1995) für eine Diskussion der Notwendigkeit sowie der Vor- und Nachteile sozialer Experimente.

$$B^{CIA} = E\{[E(Y^n | X, S=1) - g^n(X)] | S=1\}; \quad (10)$$

$$B^{CIA} = 0 \Leftrightarrow E[g^n(X) | S=1] = E(Y^n | S=1)$$

Die Annahme CIA impliziert nun sogar, daß der Bias an jeder Stelle in χ (d.h. für jede Kombination von Charakteristika, die sich in χ befindet) gleich null ist, da aus CIA $g^n(x) = E(Y^n | X=x, S=1)$ folgt, also:

$$B^{CIA}(x) = 0, \quad \forall x \in \chi; \quad \text{mit } B^{CIA}(x) = E(Y^n | X=x, S=1) - g^n(x). \quad (CIA')$$

Die Identifikation von $\theta_0^{CIA}(z)$ ist unproblematisch, wenn $z \in \chi$. Gleichung (9) modifiziert sich dann zur folgenden Gleichung (9'):²²

$$\theta_0^{CIA}(z) = g^t(z) - E[g^n(X) | Z=z, S=1]. \quad (9')$$

Damit CIA in der praktischen Anwendung plausibel ist, wird X in vielen Fällen eine große Dimension besitzen, und somit eine Schätzung von $g^n(x)$ erschweren. Jedoch haben Rosenbaum und Rubin (1983) gezeigt, daß sich mit Hilfe von sogenannten *balancing scores* das Dimensionsproblem umgehen läßt. Wenn die Funktion $b(x)$ ein *balancing score* ist, dann gilt folgendes:

$$Y^n \amalg S | X=x \Rightarrow Y^n \amalg S | b(X)=b(x); \quad \forall x \in \chi, \quad (11)$$

Hierfür muß die Funktion $b(x)$ so gewählt werden, daß sie Bedingung (12) erfüllt:

$$E[S | b(X)=b(x)] = E[S | X=x] = P(x), \quad \forall x \in \chi. \quad (12)$$

Die bedingte Teilnahmewahrscheinlichkeit $P(x)$ wird auch als *propensity score* bezeichnet. Es muß außerdem $0 < P(x) < 1$ gelten, d.h. in X darf es keine Charakteristika geben, für die die Wahrscheinlichkeit, an der Maßnahme teilzunehmen, null oder eins ist.²³ Beispiele für zu CIA kompatiblen *balancing scores* sind z.B. $b(x)=P(x)$, $b(x)=x$, oder $b(x)=[P(x), \tilde{x}]$, wobei $\tilde{x}$ einen Untervektor von x bezeichnet. Die Wahl von $b(x)=P(x)$ hat den Vorteil, daß die Dimension auf eins reduziert wird. Der Nachteil ist aber, daß $P(x)$ unbekannt ist und geschätzt werden muß.

²² Es wird in (9') erlaubt, daß z nicht nur ein Punkt, sondern eine große Untermenge von χ ist. Dies ist der Fall, wenn z über weniger Charakteristika Aussagen trifft als x . Der durchschnittliche kausale Effekt für Personen mit Charakteristika außerhalb der Menge χ ist mit CIA nicht identifizierbar.

²³ Das heißt, die Verteilungen der Teilnehmer und Nichtteilnehmer müssen sich bezüglich der relevanten Charakteristika überlappen.

Die Variablenauswahl für x spielt bei der Plausibilität von CIA eine entscheidende Rolle. Es wurde zuvor schon erwähnt, daß X alle Faktoren enthalten sollte, die Y^t und Y^n gemeinsam beeinflussen könnten. Desweiteren muß $X (= X^t S + (1-S) X^n)$ exogen sein, in dem Sinne, daß die Beobachtung nicht davon abhängt, in welchem potentiellen Zustand das Individuum sich befindet ($X=X^t=X^n$).²⁴ Daraus folgt, daß Y^t und Y^n selbst exogen sind, während alle Funktionen von $Y (= Y^t S + (1-S) Y^n)$ endogen sind. Im allgemeinen ist eine direkte Konditionierung auf Y^t und Y^n allerdings nicht möglich. Spielt jedoch der erwartete Effekt der Maßnahme bei der Teilnahmeentscheidung eine Rolle, so ist es wichtig, Faktoren zu finden, die Y^t und Y^n gut approximieren können. Diese Faktoren sollten entsprechend als Komponenten von X berücksichtigt werden. Hat man Paneldaten zur Hand, mit denen die Zeit vor und nach der Maßnahme beobachtbar ist, so läßt sich bezüglich der oben definierten Exogenität von Variablen folgendes festhalten: Zeitkonstante bzw. nicht beeinflussbare Variablen sind immer exogen. Nach Maßnahmenende datierte Variablen sind oft problematisch: Hat ihr Wert etwas mit der tatsächlichen Teilnahme zu tun, dann sind sie endogen. Oder sie haben nichts mit der Teilnahme zu tun, dann brauchen sie auch nicht in X aufgenommen zu werden. Bei zeitvariablen Variablen vor der Teilnahme ist genau zu prüfen, ob die zukünftige (evt. antizipierte) Teilnahme schon Auswirkungen auf ihre Realisation hat (z.B. langfristiges Nutzenmaximierungsverhalten der Individuen).

3.4.2 Schätzung

Hat man θ_0 identifiziert, so stellt sich die Frage nach geeigneten Schätzmethoden. Da CIA in der empirischen Praxis große Bedeutung hat, sollen im folgenden einige Schätzverfahren diskutiert werden, die in diesem Fall relevant sind. Zur Vereinfachung wird im folgenden davon ausgegangen, daß eine Zufallsstichprobe der Größe N mit N^t Teilnehmern und $(N - N^t)$ Nichtteilnehmern zur Verfügung steht. In gewissem Sinne die naheliegende Lösung ist eine Regression mit Y als abhängiger Variable und (X, S) als unabhängigen Variablen. Dieses Verfahren ist allerdings wenig attraktiv, denn zum einen hängt die konsistente Schätzung von θ_0 von der gewählten funktionalen Form der Regressionsfunktion ab, und zum anderen muß spezifiziert werden, wie θ_0 (der Regressionskoeffizient von S) über die Bevölkerung variiert (oft wird hier eine Konstanz von θ_0 unterstellt, die in vielen Fällen wenig realistisch ist).²⁵

²⁴ Siehe Fußnote 7. Diese Definition ist notwendig, um nicht auf Realisationen von Variablen zu konditionieren, die Teil der ‚Wirkung‘ der Maßnahme sind. Bei einer derartigen Konditionierung wäre $\hat{\theta}_0^{CIA}$ nicht mehr als kausaler Effekt interpretierbar, da nun das entsprechende counterfactual $E(Y^n | X=x^t, S=1)$ nicht mehr gleich dem aus den Daten identifizierbaren bedingt Erwartungswert $g(x^n) \equiv E(Y^n | X=x^n, S=1)$ [$\neq E(Y^n | X=x^t, S=1)$] wäre. X -Variablen, für die $X=X^t=X^n$ gilt, werden von Holland (1986) auch als „Attribute“ bezeichnet.

²⁵ Auch können Ausreißer die Ergebnisse stark beeinflussen, insbesondere wenn sich Teilnehmer und Kontrollbeobachtungen stark in termini von X unterscheiden. Dieses Problem wird verschärft, wenn N relativ groß im Vergleich zu N^t ist, da dann die Regressionsgerade ($x\beta$) weitgehend von der Kontrollgruppe bestimmt wird. θ_0 bezieht sich aber nur auf bedingte Erwartungswerte in der Gruppe der Teilnehmer.

Eine Alternative zu diesem einfachen Ansatz besteht darin, die beiden Komponenten von Gleichung (9) einzeln zu untersuchen. Ein naheliegendes und effizientes Schätzverfahren für $E(Y^t | S = 1)$ besteht im arithmetischen Mittel der Y in der Stichprobe der Teilnehmer:

$$\hat{E}(Y^t | S = 1) = \hat{g}^t = \frac{1}{N^t} \sum_{i=1}^{N^t} y_i. \quad (13)$$

Wenn entsprechende Momente existieren wird unter der hier getroffenen Annahme, daß die Beobachtungen als iid Ziehungen aus der entsprechenden Population zustande kamen, $\hat{E}(Y^t | S = 1)$ konsistent und asymptotisch normalverteilt sein mit Varianz $\frac{1}{N^t} \text{Var}(Y^t | S = 1)$. Da es sich um unabhängige Realisation aus der Teilnehmerpopulation handelt, kann $\text{Var}(Y^t | S = 1)$ mit der Stichprobenvarianz der Ergebnisse in der Unterstichprobe der Teilnehmer geschätzt werden ($S_{y^t}^2$). Es ist damit noch ein Schätzverfahren für $E\{g^n[b(x)] | S = 1\}$ mit $g^n[b(x)] = E[Y^n | b(X) = b(x), S = 0]$ zu finden. Es sei darauf hingewiesen, daß sich der Erwartungswert auf die Verteilung von X in der Stichprobe der Teilnehmer bezieht. Falls man ein Schätzverfahren $\hat{g}^n[b(x)]$ erhalten hat, das zumindest für eine große Kontrollgruppe $g^n[b(x)]$ nahezu unverzerrt schätzt, so ist die Schätzung des äußeren Erwartungswerts problemlos und kann z.B. folgendermaßen erfolgen:

$$\hat{E}_X[g^n(X) | S = 1] = \frac{1}{N^t} \sum_{i=1}^{N^t} \hat{g}^n[b(x_i)]. \quad (14)$$

Da die Schätzung der beiden Komponenten von θ_0 mit y_i verschiedener Personen erfolgt, ergibt sich $\hat{\theta}_N^{CIA}$ und dessen Varianz bei einer Stichprobengröße N als:

$$\hat{\theta}_N^{CIA} = \frac{1}{N^t} \sum_{i=1}^{N^t} \{y_i - \hat{g}^n[b(x_i)]\}; \quad (15)$$

$$\text{Var}(\hat{\theta}_N^{CIA}) = \frac{1}{N^t} \left\{ S_{y^t}^2 + \frac{1}{N^t} \text{Var} \left[\sum_{i=1}^{N^t} \hat{g}^n[b(x_i)] \right] \right\}. \quad (16)$$

Unter üblichen Bedingungen ist $\hat{\theta}_N^{CIA}$ konsistent für θ_0 und $\sqrt{N^t}(\hat{\theta}_N^{CIA} - \theta_0)$ ist asymptotisch normalverteilt mit Varianz $N^t \text{Var}(\hat{\theta}_N^{CIA})$.²⁶ Somit reduziert sich das Problem darauf, eine geeignete Schätzung $\hat{g}[b(x_i)]$ zu finden. Die einfachste Methode besteht wohl darin, eine lineare oder auch nichtlineare, parametrische Regression (OLS, Probit, Tobit etc.) in der

²⁶ Siehe Heckman et al. (1996) für eine umfassende und vollständige Diskussion derartiger Verfahren und deren Eigenschaften.

Kontrollgruppe mit einem beliebig gewählten $b(x)$ (z.B. $b(x) = x$) durchzuführen und dann diese Regressionsfunktion an der Stelle x_i zu berechnen. Da eine Fehlspezifikation der Regressionsfunktion allerdings zu inkonsistenter Schätzung von θ_0 führt, sind nichtparametrische Regressionsverfahren im allgemeinen vorzuziehen (Härdle, 1990).²⁷ Um Probleme mit dem „Fluch der Dimensionalität“ zu vermeiden, empfiehlt es sich, $b(x) = P(x)$ zu wählen. Methoden für nichtparametrische Regressionen mit nur einem (nahezu) kontinuierlichen Regressor sind gut anwendbar und in vielen Softwarepaketen enthalten.

Ein noch einfacheres nichtparametrisches Verfahren ist *matching* (Rosenbaum, Rubin, 1983, 1985). Die Idee besteht darin, für jeden Teilnehmer i (mindestens) einen Nichtteilnehmer j zu finden der (annähernd) die gleichen Charakteristika $b(x_i)$ besitzt. Dessen beobachtetes $y_{j(i)}^n = y_{j(i)}$ ($j \in \{N^t + 1, \dots, N\}$) ersetzt dann die Funktion $\hat{g}^n[b(x_i)]$ in den Gleichungen (15) und (16).²⁸ Wird ein Matching-Algorithmus verwendet, der jede Kontrollbeobachtung nur einmal verwendet, dann vereinfacht sich auch die Varianzberechnung stark und die Schätzung verläuft analog zur Schätzung der Varianz für die Teilnehmer. Außerdem kann die Methodik des Matching dazu verwendet werden, aus der Untergruppe der Nichtteilnehmer eine Stichprobe zu erhalten, die dann wie eine Kontrollgruppe aus einem sozialen Experiment verwendet werden kann. Zum Beispiel wird diese Eigenschaft von Hujer et al. (1997 a) ausgenutzt, um ein von Ham, Lalonde (1996) für Experimente vorgeschlagenes Verfahren zur Schätzung von Verweildauermodellen (zur Vermeidung von Selektionsproblemen durch Zensierung) zur Evaluierung von Trainingsprogrammen auf eine westdeutsche „gematchte“ Stichprobe anzuwenden. Der Vorteil der nichtparametrischen Verfahren besteht in ihrer Robustheit gegenüber Fehlspezifikation des bedingten Erwartungswertes $g^n(x)$, bzw. $g^n[b(x)]$. Der Nachteil ist, daß $b(x) = x$ als *balancing score* aufgrund der oft großen Dimensionen von X nicht in Frage kommt, sondern stattdessen $P(x)$ geschätzt werden muß. Üblich ist hier die Schätzung von Logit- oder Probitmodellen. Um aber nicht „durch die Hintertür“ einer Fehlspezifikation von $P(x)$ eine Inkonsistenz von θ_0 zu erhalten, sollte die Spezifikation von $P(x)$ intensiv getestet werden.²⁹

Für beide nichtparametrischen Methoden ist es essentiell, daß der Wertebereich von X für die Populationen $S=1$ und $S=0$ vollständig überlappt, da sonst für einige Teilnehmer keine vergleichbaren Nichtteilnehmer gefunden werden können. Diese Teilnehmer sollten daher, von der Analyse ausgeschlossen werden. Dies hat allerdings den Nachteil, daß der geschätzte Effekte nur noch für eine Untergruppe der Teilnehmer relevant ist. Dieses Problem besteht

²⁷ Wenn sich die Verteilungen $(X|S=1)$ und $(X|S=0)$ stark unterscheiden, dann sind Fehlspezifikationen bei Regressionsmethoden besonders problematisch, da die Lage der Regressionskurve in $X|S=0$ bestimmt und dann in den Raum $X|S=1$ projiziert wird.

²⁸ Die Anwendung von *Matching*-Verfahren läßt sich zumindest bis auf Fechner (1860, insb. S.76 ff., 89 ff. und 113 ff.) zurückverfolgen. Ich danke James Heckman für den Hinweis auf diese Quelle.

²⁹ Leider sind dem Autor keine Arbeiten anderer Autoren bekannt, in denen dies intensiv durchgeführt wurde (vgl. Lechner, 1995 a, 1996 a, 1996 b).

nicht bei parametrischen Methoden, da man aufgrund des geschätzten Modells immer in den Bereich fehlender Daten projizieren kann.

Vergleicht man die nichtparametrischen Methoden miteinander, so ist ein Trade-off zwischen Einfachheit und Effizienz festzuhalten. Die nichtparametrische Regression ist effizienter, da für die Schätzung eines jeden $\hat{g}^n(x_i)$ mehrere Kontrollbeobachtungen verwendet werden, während bei Matching – zumindest in seiner einfachsten Form – nur eine einzige Kontrollbeobachtung pro Teilnehmer verwendet wird. Beim Matching ist aber dafür zumindest die Berechnung der Varianz erheblich einfacher. Ein Teil der Einfachheit des Matching geht allerdings verloren, wenn sich keine exakten Matches in den Daten finden lassen. Um hier keinen systematischen Bias entstehen zu lassen, sollten die Ergebnisse des Matching nachträglich z.B. mit lokalen Regressionsmethoden korrigiert werden (siehe Lechner, 1995 a).³⁰ Da die nichtparametrische Regression lokale Mittelwerte um x_i bildet, ist eine derartig Korrektur bei dieser Methode nicht notwendig.

Ein Problem, von dem alle Verfahren betroffen sind, tritt auf, wenn die Teilnahme für die einzelnen Individuen zu unterschiedlichen Zeitpunkten beginnt. Falls dann für X Variablen relevant sind, die sich auf einen Zeitabstand vor der Teilnahme beziehen, wie z.B. der Arbeitsmarktstatus im Monat vor der Teilnahme, dann sind diese a priori für die Kontrollgruppe unbekannt (bzw. variieren, je nachdem auf welches Anfangsdatum man sich bezieht). $P(x)$ kann daher nicht geschätzt werden und die Reduktion der Dimension ist schwierig. Nichtparametrische Regressionsmethoden sind dann schlecht oder gar nicht anwendbar. Lechner (1995 a) hat jedoch mehrere Matching-Verfahren vorgeschlagen, die zwar etwas komplizierter als die vorgestellte Variante sind, aber anwendbar bleiben. Eines dieser Verfahren wurde inzwischen in mehreren Studien recht erfolgversprechend eingesetzt (Hujer et al., 1997 a, Lechner, 1996 a, b).

3.5 Symmetrische Verzerrung der Schätzungen vor und nach der Maßnahme

Es kann in der jeweiligen Anwendung aufgrund unzureichender Daten und/oder unzureichenden Wissens über den Selektion der Teilnehmer in die Maßnahme, plausibel sein, daß weder CIA noch CIA' Gültigkeit besitzen. Dann kommen im Prinzip nur noch drei Möglichkeiten der Identifikation von θ_0 in Betracht. Falls eine Variable im Datensatz vorhanden ist, die die Eigenschaften einer „Instrumentalvariablen“ erfüllt, dann sind entsprechende nichtparametrische IV-Verfahren relevant (vgl. Angrist et al., 1996, und Heckman, 1997). Leider ist es aufgrund der notwendigen Kürze dieses Aufsatzes nicht möglich, auf diese Verfahren näher einzugehen. Es sei jedoch trotzdem auf zwei gravierende Nachteile hingewiesen. Zum einen gibt es in „realen Situationen“ nur selten Variablen, die die in dieser Literatur verlangten strengen Anforderungen an ein Instrument erfüllen. Zum

³⁰ Heckman et al. (1996 b) zeigen, daß Matching-Schätzungen erheblich verzerrt sein können, wenn CIA oder die Bedingung sich überlappenden Wertebereiche nicht erfüllt sind.

anderen ist es nicht möglich, $\theta_0(z)$ zu identifizieren. Es ist nur möglich $\theta_0(z)$ für eine besondere (unbeobachtbare) Gruppe zu identifizieren, für die das gewählte Instrument eine bestimmte Wirkung besitzt (sogenannte *local average treatment effects*, *LATE*). Die zweite Alternative besteht darin, ein parametrisches oder semiparametrisches Modell zu spezifizieren und somit die fehlende Information (und in vielen Fällen nicht nur diese!) durch statistische Annahmen zu ersetzen. Dieser Ansatz wird in Abschnitt 3.6 kurz behandelt.

Die dritte Möglichkeit besteht darin anzunehmen, daß die Verzerrung nicht völlig zufällig ist, sondern eine gewissen Systematik (Struktur) aufweist. Im folgenden soll, ausgehend von Heckman et al. (1996 a), eine nichtparametrische Methode vorgestellt werden, die in solchen Fällen sehr interessant für die empirische Anwendung erscheint. Die Grundidee ist die Annahme, daß die Verzerrung der Schätzung, die durch die Verletzung von CIA entsteht, zu mindestens einem beliebigen Zeitpunkt *vor* (τ , $\tau < 0$) und *nach* (t , $t > 0$) der Maßnahme gleich ist. Somit kann sie durch geschickte Differenzbildung eliminiert werden. Ein Vorteil der nichtparametrischen Variante liegt darin, daß CIA (oft) als ein Spezialfall davon angesehen werden kann.³¹ Ist die Annahme CIA nicht erfüllt, also $B(x) \neq 0$, schlagen Heckman et al. (1996 a) folgende Bedingung als identifizierende Annahme vor:³²

$$B_t^{CIA}(x_t) - B_\tau^{CIA}(x_\tau) = 0 \quad \text{für mindestens ein } \tau < 0 \text{ und } t > 0, x \in \mathcal{X}. \quad (17)$$

Der mit dieser Annahme einer zeitstabilen Verzerrung (*Bias-Stability-Assumption*, BSA) identifizierbare Effekt, ergibt sich als:³³

$$\theta_{0,t}^{BSA} = \theta_{0,t}^{CIA} - \theta_{0,\tau}^{CIA}. \quad (18)$$

Die Verzerrung erhält man nach einigen Umformungen als:

$$B_t^{BSA} = B_t^{CIA} - B_\tau^{CIA} - \theta_{0,\tau}. \quad (19)$$

Dadurch wird deutlich, daß (17) weder hinreichend noch notwendig für die Konsistenz von $\theta_{0,t}^{BSA}$ ist. Die hinreichende Bedingung lautet stattdessen:

$$B_t^{BSA} = 0 \Leftrightarrow E(Y_t^n | S = 1) - E_X[g_t^n(X) | S = 1] = E_X\{[g_t^t(X) - g_\tau^n(X)] | S = 1\} \\ \text{für mindestens ein } \tau, t, t \neq \tau, x \in \mathcal{X}. \quad (\text{BSA})$$

³¹ Siehe Meyer (1995) für ähnliche Methoden im Regressionsmodell.

³² Alle Symbole, die im folgenden mit „ τ “ oder „ t “ indiziert sind, beziehen sich auf den jeweiligen Zeitpunkt.

³³ Man könnte Schätzverfahren die auf BSA zur Identifikation beruhen auch als (nichtparametrische) Differenzen in Differenzen-Verfahren (*Differences in Differences*) bezeichnen. Die zweite Differenz steht für ist die Differenz zwischen Teilnehmergruppen und gematchten Kontrollgruppen.

Diese Bedingung ist weniger restriktiv, als die von Heckman et al. (1996 a) vorgeschlagene und in Gleichung (17) enthaltene Bedingung. Ökonomisch am einfachsten zu interpretieren ist sie für den Fall $\theta_{0,\tau} = 0$. Dieser Fall ist zum Beispiel dann gegeben, wenn davon ausgegangen werden kann, daß die spätere Teilnahme zu einem bestimmten Zeitpunkt vor Beginn ($\tau < 0$) noch keine Wirkung hat.³⁴ In diesem Fall impliziert (BSA), daß die durch CIA entstehende Selektionsverzerrung, also $E[Y^n | S = 1] - E[g^n(X) | S = 1]$, zu den beiden Zeitpunkten t und τ identisch ist. Diese Annahme erscheint in der Praxis für viele Fälle relevant zu sein.

Es ist interessant, θ_0^{BSA} und θ_0^{CIA} für einige Fälle zu vergleichen. Betrachten wir zuerst den Fall, daß $X_t = X_\tau$ gilt und zur Schätzung von $\hat{\theta}_N^{CIA}$ einfache Matching-Methoden verwendet wurden. In diesem Fall bietet es sich an, ein Matchingverfahren zu verwenden, mit dem für $\hat{\theta}_N^{CIA,t}$ und $\hat{\theta}_N^{CIA,\tau}$ die gleichen Kontrollbeobachtungen gematcht werden.

Dann hat $\hat{\theta}_N^{BSA}$ die einfache und intuitiv sehr einleuchtende Form:

$$\hat{\theta}_N^{BSA} = \frac{1}{N} \sum_{i=1}^{N^t} [(y_{i,t}^t - y_{i,\tau}^t) - (y_{j(i),t}^n - y_{j(i),\tau}^n)] .$$

Der Fall $x_t = x_\tau$ ist allerdings wenig sinnvoll, wenn $Y_t \in X$, denn daraus folgt direkt $\theta_{0,\tau}^{BSA} = 0$ und $\theta_{0,t}^{BSA} = \theta_{0,t}^{CIA}$. Somit ist nichts gegenüber CIA gewonnen. Relevanter ist beispielsweise der Fall, bei dem aufgrund eines nicht hinreichend informativen Datensatzes $\{..., Y_3, Y_1, Y_{-1}, Y_{-3}, ...\}$ nicht beobachtbar sind, jedoch $Y_t \perp\!\!\!\perp S | Y_{-1} (t > 0)$, aber nicht $Y_t \perp\!\!\!\perp S | Y_{-2} (t > 0)$ gilt. Wenn dann BSA plausibel erscheint, kann man z.B. $\theta_{N,2}^{CIA}$ mit Y_{-2} als Attribut und $\theta_{N,-2}^{CIA}$ mit Y_{-4} als Attribut schätzen (z.B. mittels Matching) und $\hat{\theta}_{N,2}^{BSA} = \hat{\theta}_{N,2}^{CIA} - \hat{\theta}_{N,-2}^{CIA}$ als konsistenter Schätzer für $\theta_{0,2}$ erhalten. Die weitere Diskussion von BSA, und insbesondere des Zusammenhangs zu CIA, sowie die „optimale“ Wahl von t , τ , x_t und x_τ geht über das Ziel dieser Arbeit hinaus und wird zukünftigen Arbeiten vorbehalten bleiben. Jedoch erscheint die Methode aufgrund ihrer Einfachheit und Flexibilität a priori sehr attraktiv für empirische Anwendungen.³⁵

3.6 Identifikation mit parametrischen Modellen

Die bisherigen Ansätze versuchten, Identifikation von θ_0 zu erreichen, indem institutionelles Wissen über den Teilnahmeprozess mit hinreichend informativen Daten verknüpft wurde. Dabei wird auf ökonomisch oft schwer nachvollziehbare Annahmen über funktionale Formen von Erwartungswerten oder auf sonstige Verteilungsannahmen verzichtet. Der Ansatz, der nun besprochen werden soll, basiert hingegen auf einer mehr oder weniger vollständigen, d.h. parametrischen oder semiparametrischen, Modellierung sowohl des Selektionsprozesses zur

³⁴ Dies ist im übrigen auch eine notwendige Annahme für eine sinnvolle Interpretation des von Heckman und Hotz (1989) vorgeschlagenen sogenannten Pre-Programm-Tests.

³⁵ Dies gilt insbesondere, da BSA im Prinzip nicht nur mit Paneldaten, sondern auch mit wiederholten Querschnitten anwendbar ist.

Teilnahme als auch der Prozesse, denen die Ergebnisvariablen unterliegen. Dabei werden ein struktureller und ein reduzierter-Form Ansatz unterschieden.

Mit dem strukturellen Ansatz wird beispielsweise ein individuelles Nutzenmaximierungskalkül postuliert, das die Teilnahmebereitschaft steuert. Ein solches Modell enthält oft auch eine Spezifikation der Abhängigkeit der Ergebnisvariablen Y_t von der Teilnahme. Schließt man ein derartiges Modell mit geeigneten Annahmen über die gemeinsame Verteilung von dem Modell hinzugefügten Fehlertermen, so ist es möglich, θ_0 zu identifizieren. Mit den Daten können dann die Prognosen des Modells einschließlich der rein statistischen Annahmen überprüft werden. Verzichtet man alternativ auf alle Annahmen, die nicht ökonomisch motiviert sind, verliert man zwar die Identifikation von θ_0 , kann θ_0 aber analog zu Manski (1990, siehe auch Abschnitt 3.1) mit den Implikationen des Modells auf ein mehr oder weniger großes Intervall einschränken (vgl. Rettore, 1997).

Ein in der empirischen Literatur häufiger verwendeter Ansatz soll hier als reduzierter-Form-Ansatz (RFA) bezeichnet werden. Beim RFA wird die Rolle der Theorie typischerweise auf die Auswahl von notwendigen Gleichungen und Variablen beschränkt. Die eigentliche Motivation dieses Ansatzes besteht darin, daß man mit einer Modellierung bzw. der Verwendung semiparametrischer Selektionsmodelle auch konsistente Schätzungen für θ_0 erhalten kann, wenn CIA nicht gilt. Die sehr technische Literatur über die Identifikation und Schätzung semiparametrischer Selektionsmodelle zeigt, daß (an Stelle von CIA oder BSA) auch mit diesen Modellen Annahmen getroffen werden müssen, deren Gültigkeit mit ökonomischen Überlegungen kaum bis gar nicht nachzuvollziehen ist (vgl. für einen knappen Überblick Heckman et al., 1996 a).³⁶

In der empirischen Anwendung für die BRD sind solche semiparametrischen Verfahren allerdings noch nicht verwendet worden (was aufgrund vieler Komplikationen bei der Anwendung auch verständlich ist). Die zur Zeit nach meiner Meinung in der deutschen Evaluationsliteratur am weitesten fortgeschrittenen Arbeiten, die dem Modell-Evaluationstyp angehören, sind die Arbeiten von Fitzenberger und Prey (1996, 1997) und Prey (1997).³⁷ Daher möchte ich anhand dieser Arbeiten, die im folgenden mit FPP abgekürzt werden sollen, kurz die Struktur des Vorgehens erläutern und die Probleme aufzeigen.³⁸

Gehen wir dabei der Einfachheit halber nur von der Ergebnisvariable Arbeitslosigkeit (AL) aus. FPP spezifizieren nun drei Gleichungen: Eine dynamische Probitgleichung für die AL, eine Probitgleichung für die Trainingsteilnahme und eine nur von zeitkonstanten Regressoren abhängige Probitgleichung für die Panelmortalität, die in dem verwendeten Datensatz ein

³⁶ Für die Spezialisten unter den Lesern sei nur an Bedingungen wie „Index-Suffizienz“, „Ausschlußrestriktionen“ und „Identifikation im Unendlichen“ erinnert. Wenn CIA oder BSA gültig ist, sind derartige Annahmen nicht notwendig.

³⁷ Fitzenberger und Prey (1996, 1997) verwenden einen Datensatz, der nicht informativ genug ist, um CIA plausibel zu machen. Auf Prey (1997) trifft dieses Argument allerdings nicht notwendigerweise zu.

³⁸ Vgl. auch die Arbeit von Fitzenberger und Prey in diesem Band. Eine Diskussion verschiedener verwandter Modelltypen findet sich beispielsweise auch in Lechner (1995 b).

besonderes Problem ist. Der Effekt von Training ist modelliert zum einen als Dummy-Variable, die nach der Teilnahme gleich eins ist, und zum anderen als Dummy-Variable, die vor der Teilnahme gleich eins ist. Letztere dient im Sinne von Heckman und Hotz (1989) als „Pre-Programm-Test“, d.h. es wird $\theta_{0,\tau} = 0$ ($\tau < 0$) unterstellt und von einer konsistenten Schätzung des entsprechenden Koeffizienten ausgegangen.³⁹ Zwischen θ_0, γ_0 und ξ_0 wird nicht explizit unterschieden, d.h. es werden im Prinzip (bis auf mögliche Effekte der Selektionsterme) homogene Effekte unterstellt. Das Modell wird durch Hinzufügen von unbeobachteten Störtermen geschlossen, für die eine multivariate Normalverteilungsannahme getroffen wird. Zur einfacheren Berechenbarkeit wird für die gesamte Kovarianzmatrix (über Zeit und Gleichungen) eine „Ein-Faktor“-Struktur unterstellt.

Zur Schätzung wird die Maximum-Likelihood-Methode verwendet. Dabei kann aufgrund der getroffenen Restriktionen auf die Verwendung aufwendiger Simulationsmethoden verzichtet werden. Die Evaluationsergebnisse unterscheiden sich dann teilweise mehr, teilweise weniger von den Ergebnissen, die Lechner (1995 a, 1996 a, b) und Hujer et al. (1997 a) mit nichtparametrischen Methoden erzielt haben.

Dieses Beispiel weist meiner Ansicht nach darauf hin, welche Probleme allgemein bei Anwendung eines derartigen Ansatzes anzutreffen sind. Zum einen ist ein erheblicher Spezifikations-, Programmier- und Rechenaufwand notwendig, um die Parameter derartiger Modelle mit einer auch nur halbwegs flexiblen Parametrisierung schätzen zu können. Zum anderen ist es aufgrund der Handhabbarkeit notwendig, Annahmen zu treffen, von deren Inkorrektheit man a priori ausgehen kann (Homogenität der Effekte, Normalverteilungen, etc.).⁴⁰ Die Frage ist aber natürlich nicht, ob das ökonometrische Modell richtig oder falsch ist – ökonometrische Modelle und Annahmen sind immer falsch – sondern, ob die Realität hinreichend gut abgebildet wird. Hierüber kann man sehr verschiedener Meinung sein, aber diese Meinungen leider nur schlecht mit ökonomischen Überlegungen zu begründen. Es ist inzwischen meine Überzeugung, daß sich die approximative Validität von Annahmen wie BAC, IA, CIA, und sogar BSA besser überprüfen läßt – da sie auf Verhaltensannahmen basieren – als rein technische Annahmen über unbekannte Störterme und funktionale Formen. Dies gilt auch für die Abschätzung der Folgen, die zu erwarten sind, wenn derartige Annahmen verletzt werden. Es muß jedoch betont werden, daß es sich hier um eine subjektive Einschätzung handelt, die durch keinen mathematischen Beweis bestätigt werden kann. Auch sollte ich zum Abschluß hervorheben, daß ich die Arbeiten von FPP als Beispiel gewählt habe, weil ich sie als hervorragende Arbeiten innerhalb dieses Ansatzes der RFA verstehe.

³⁹ Letzteres ist aber insbesondere fraglich, da die Likelihood-Funktion die „Endogenität“ dieser Variable nicht berücksichtigt und der „Pre-Programm-Test“ auf Fehlspezifikation hindeutet. Dieser Punkt ist relevant, da die Interpretation der Ergebnisse entscheidend von dieser Variablen abhängt.

⁴⁰ Darüberhinaus ist die Schätzung der Pre-Test-Dummies möglicherweise inkonsistent. Dabei handelt es sich aber weniger um ein spezifisches Problem dieses Ansatzes, sondern vielmehr um ein allgemeines Problem, wie es entstehen kann, wenn man einen Fehlspezifikationstest zur Schätzung unter der Annahme der Fehlspezifikation verwendet.

4 Abschließende Bemerkungen

In diesem Aufsatz wurden wichtige Methoden zur Identifikation von kausalen Effekten von staatlichen Eingriffen am Beispiel von Fortbildungs- und Umschulungsprogrammen (Training) vorgestellt und deren Implikationen, sowie auch teilweise die Eigenschaften jeweils in Frage kommender Schätzmethoden diskutiert. Welche Annahmen und welche Schätzverfahren sind nun vorzuziehen? Auf diese Frage gibt es leider keine allgemeingültige Antwort. Es kommt immer auf die jeweiligen Umstände, das Wissen des Untersuchenden und die ihm zur Verfügung stehenden Daten an. Oder um es mit James Heckman und Koautoren zu sagen: "Thus the choice of an appropriate econometric model critically depends on the properties of the data on which it is applied. ..." (Heckman et al., 1996 a, S.11). Auch wenn dieses Zitat impliziert, daß jede ernstzunehmende Evaluationsstudie auch in Zukunft einen gewissen Aufwand erfordern wird („Evaluationen am Fließband“ wird es selten geben), so muß doch für Deutschland gelten, daß dringend mehr ökonometrische Evaluationsstudien notwendig sind, um die hohen Ausgaben zum Beispiel für die aktive Arbeitsmarktpolitik besser lenken zu können.⁴¹ Wesentliche Voraussetzungen hierfür sind unter anderem die Verfügbarkeit von besseren Daten und ausreichenden Forschungsressourcen. Daher soll dieser Aufsatz mit dem Aufruf enden, genau diese Ressourcen der Wissenschaft zur Verfügung zu stellen.

Literaturverzeichnis

- Angrist, J. D., Imbens, G. W., Rubin, D. B. (1996): "Identification of Causal Effects Using Instrumental Variables", in: *Journal of the American Statistical Association*, 91, 444-472, mit Diskussion von by J.J. Heckman, R.A. Moffitt, J.M. Robins und S. Greenland, und R.P. Rosenbaum.
- Ashenfelter, O. (1978): "Estimating the Effect of Training Programs on Earnings ", *The Review of Economics and Statistics*, 58, 47-57.
- Bloom, H. S., Orr, L. L., Bell, S.H., Cave, G., Doolittle, F., Lin, W., Bros, J. M. (1997): "The Benefits and Costs of JTPA Title II-A Programs", in: *Journal of Human Resources*, 32, 549-576.
- Burtless, G. (1995): "The Case for Randomized Field Trials in Economic and Policy Research", in: *Journal of Economic Perspectives*, 9, 63-84.
- Card, D., Sullivan, D. (1988): "Measuring the Effect of Subsidized Training Programs on Movements in and out of Employment", in: *Econometrica*, 56, 497-530.
- Clements, N., Heckman, J. J., Smith, J.A. (1994): "Making the most out of social experiments: Reducing the intrinsic uncertainty in evidence from randomized trials with an application to the National JTPA experiment", in: *NBER Technical Working Paper*, 166.
- Eichler, M., Lechner, M. (1996): "Public Sector Sponsored Continuous Vocational Training in East Germany: Institutional Arrangements, Participants, and Results of Empirical Evaluations", in: *Beiträge zur angewandten Wirtschaftsforschung*, Universität Mannheim, 549-96.
- Fechner, G. T. (1860): *Elemente der Psychophysik*, Leipzig: Breitkopf und Hartel, (3. unveränderte Auflage, 1907).
- Fitzenberger, B., Prey, H. (1996): "Training in East Germany: An Evaluation of the Effects on Employment and Earnings", *Unpublished manuscript*, University of Konstanz.
- Fitzenberger, B., Prey, H. (1997): "Assessing the Impact of Training on Employment: The Case of East Germany", in: *IFO Studien*, 43(1), 71-116.

⁴¹ Das große Wissen über die Trainings- und Arbeitsmarktprogramme der USA hilft nur wenig, da die Programme in der BRD in einem institutionell wesentlich anderen Arbeitsmarkt stattfinden.

- Fitzenberger, B., Prey, H. (1998): "Beschäftigungs- und Verdienstwirkungen von Weiterbildungsmaßnahmen im ostdeutschen Transformationsprozeß: Eine Methodenkritik", dieser Band.
- Garfunkel, I., Manski, C. F., Michalopoulos, C. (1992): "Micro Experiments and Macro Effects", in: C.F. Manski and I. Garfunkel (Hrsg.), in: *Evaluating Welfare and Training Programs*, Cambridge: Harvard University Press, 253-273.
- Greenberg, D. H. (1997): "The Leisure Bias in Cost-Benefit Analyses of Employment and Training", in: *Journal of Human Resources*, 32, 413-439.
- Gritz, R.M. (1993): "The impact of training on the frequency and duration of employment", in: *Journal of Econometrics*, 57, 21-51.
- Ham, J. C., Lalonde, R. J. (1996): "The Effect of Sample Selection and Initial Conditions in Duration Models: Evidence from Experimental Data on Training", in: *Econometrica*, 64, 175-205.
- Härdle, W. (1990): *Applied Nonparametric Regression*, Cambridge: Cambridge University Press.
- Heckman, J. J. (1997): "Instrumental Variables", in: *Journal of Human Resources*, 32, 441-462.
- Heckman, J. J., Ichimura, H., Smith, J. A., Todd, P. (1996 a): "Characterisation of Selection Bias Using Experimental Data", *mimeo*, October 1996.
- Heckman, J. J., Ichimura, H., Smith, J. A., Todd, P. (1996 b): "Sources of selection bias in evaluation social programs: An interpretation of conventional measures and evidence on the effectiveness of matching as a program evaluation method", in: *Proc. Natl. Acad. Sci., USA*, 13416-13420, November 1996.
- Heckman, J. J., Lochner, L., Smith, J. A., Taber, C. (1997): "The Effects of Government Policies on Human Capital Investment and Earnings Inequality", *mimeo*, June 1997.
- Heckman, J. J., Robb, R. (1985): "Alternative Methods of Evaluating the Impact of Interventions", in: J.J. Heckman, B. Singer (Hrsg.), *Longitudinal Analysis of Labour Market Data*, New York: Cambridge University Press.
- Heckman, J. J., Smith, J. A. (1995): "Assessing the Case for Social Experiments", in: *Journal of Economic Perspectives*, 9, 85-110.
- Heckman, J. J., Hotz, V. J. (1989): "Choosing Among Alternative Nonexperimental Methods for Estimating the Impact of Social Programs: The Case of Manpower Training", in: *Journal of the American Statistical Association*, 84, 862-880.
- Holland, P. W. (1986): "Statistics and Causal Inference", in: *Journal of the American Statistical Association*, 81, 945-970, mit Diskussion.
- Hübler, O. (1997): "Evaluation beschäftigungspolitischer Maßnahmen in Ostdeutschland", in: *Jahrbücher für Nationalökonomie und Statistik*, 216, 21-44.
- Hujer, R., Maurer, K., Wellner, M. (1996): "The Impact of Training on Employment: A Survey of Microeconomic Studies", in: *Discussion papers in Economics # 69*, University of Frankfurt.
- Hujer, R., Maurer, K., Wellner, M. (1997 a): "The Impact of Training on Unemployment Duration in West Germany – Combining a Discrete Hazard-Rate Model with Matching Techniques", in: *Discussion papers in Economics # 74*, University of Frankfurt.
- Hujer, R., Maurer, K., Wellner, M. (1997 b): "Estimating the Effects of Training on Unemployment Duration in West Germany – A Discrete Hazard-Rate Model with Instrumental Variables", in: *Discussion papers in Economics # 73*, University of Frankfurt.
- LaLonde, R. J. (1995): "The Promise of Public Sector-Sponsored Training Programs", in: *Journal of Economic Perspectives*, 9, 149-168.
- Lechner, M. (1995 a): "Effects of Continuous Off-the-job Training in East Germany after Unification", in: *Discussion Paper, # 95-27*, Zentrum für Europäische Wirtschaftsforschung, Mannheim, revidierte Version September 1997.
- Lechner, M. (1995 b): *Ökonometrische Modelle für qualitative Paneldaten: Das Beispiel der selbständigen Erwerbstätigkeit in den neuen Bundesländern*, Frankfurt: Campus, 1995.
- Lechner, M. (1996 a): "An Evaluation of Public Sector Sponsored Continuous Vocational Training Programs in East Germany", in: *Beiträge zur angewandten Wirtschaftsforschung # 539-96*, Universität Mannheim.
- Lechner, M. (1996 b): "Effects of Enterprise-related Continuous Vocational Training in East Germany on Individual Employment and Earnings", in: *Beiträge zur angewandten Wirtschaftsforschung # 542-96*, Universität Mannheim.

- Lechner, M. (1996 c): "Nonparametric Bounds on Employment and Income Effects of Continuous Vocational Training in East Germany", in: *Discussion Paper # 96-31*, Zentrum für Europäische Wirtschaftsforschung.
- Manski, C. F. (1990): "Nonparametric Bounds on Treatment Effects", in: *American Economic Review, Papers and Proceedings*, 80, 319-323.
- Manski, C. F., Sandefur, G. D., McLanahan, S., Powers, D. (1992): "Alternative Estimates of the Effect of Family Structure During Adolescence on High School Graduation", in: *Journal of the American Statistical Association*, 87, 25-37.
- Meyer, B. M. (1995): "Natural and Quasi-Experiments in Economics", in: *Journal of Business & Economic Statistics*, 13, 151-161.
- Prey, H. (1997): "Beschäftigungswirkungen von öffentlich geförderten Qualifizierungsmassnahmen. Eine Paneluntersuchung für Westdeutschland", in: *Discussion paper # 41*, University of Konstanz.
- Rettore, E. (1997): "The Impact of Training Programs on Duration of the First Job Search Spell", *unpublished manuscript*, University of Padova.
- Rosenbaum, P. R., Rubin, D. B. (1983): "The Central Role of the Propensity Score in Observational Studies for Causal Effects", in: *Biometrika*, 70, 41-50.
- Rosenbaum, P. R., Rubin, D. B. (1985): "Constructing a Control Group Using Multivariate Matched Sampling Methods That Incorporate the Propensity Score", in: *The American Statistician*, 39, 33-38.
- Roy, A. D. (1951): "Some Thoughts on the Distribution of Earnings", in: *Oxford Economic Papers*, 3, 135-146.
- Rubin, D. B. (1974): "Estimating Causal Effects of Treatments in Randomized and Nonrandomized Studies", in: *Journal of Educational Psychology*, 66, 688-701.
- Schnellhaaß, H., Schubert, A. (1992): "Internationale Entwicklungen der Evaluierungsmethoden für arbeitsmarktpolitische Programme", in: *MittAB*, 3/92, 371-381.
- Sobel, M. E. (1994): "Causal Inference in the Social and Behavioral Sciences", in: G. Arminger, C. C. Clogg und M.E. Sobel (Hrsg.): *Handbook of Statistical Modeling for the Social and Behavioral Sciences*, New York: Plenum Press.