Insgesamt ergeben sich für die analysierten Gruppen keine wesentlichen Unterschiede in den Ergebnissen. Für alle wird im Fall der Teilnahme an einer ABM/SAM kurz- und mittelfristig eine deutlich geringere Abgangswahrscheinlichkeit aus Arbeitslosigkeit festgestellt. Auch langfristig scheint die Teilnahme die Beschäftigungswahrscheinlichkeit zu verringern.

Vor dem Hintergrund dieser Ergebnisse ist die aktuelle Diskussion um die Zweckmäßigkeit von Arbeitsbeschaffungsmaßnahmen zu begrüßen und insbesondere die Forderung nach einer kritischen Überprüfung dieses Instruments zu bekräftigen.

Eva.Reinowski@iwh-halle.de
Birgit.Schultz@iwh-halle.de
Jürgen.Wiemers@iwh-halle.de

Evaluation arbeitsmarktpolitischer Maßnahmen: Fallstricke und Lösungsansätze

Die Maßnahmen der aktiven Arbeitsmarktpolitik sind ein kostspieliges Instrument zur Beseitigung der Arbeitslosigkeit. Allein im vergangenen Jahr wurden von der Bundesanstalt für Arbeit 19,5 Mrd. Euro dafür ausgegeben.[21] *Angesichts der angespannten Haushaltslage öffentlicher Kassen ist es daher unerlässlich, ihre Wirkung auf die Beschäftigungschancen der geförderten Arbeitslosen zu überprüfen.*

Der Erfolg solcher Maßnahmen wird oft vorschnell anhand einzelner statistischer Daten beurteilt, die für sich genommen allerdings wenig aussagefähig sind. So ist beispielsweise eine hohe Übergangsquote von Teilnehmern an einer Weiterbildungsmaßnahme in Erwerbstätigkeit während oder nach Beendigung dieser Maßnahme per se noch kein Erfolg. Vielmehr ist es möglich, dass sie auch ohne Maßnahme eine Beschäftigung gefunden hätten. Die Frage ist also, womit die Übergangsrate verglichen werden kann, um tatsächliche Erfolge festzustellen.

Der Lösung dieses Problems wird bei der wissenschaftlichen Evaluation arbeitsmarktpolitischer Maßnahmen ein hoher Stellenwert eingeräumt. In Abhängigkeit von unterschiedlichen Datenstrukturen und Fragestellungen können verschiedene Lösungswege beschritten werden.

In diesem Beitrag soll ein Einblick in die Vielfalt der angewendeten Methoden vermittelt werden. Für ausgewählte Verfahren werden dazu deren Grundannahmen sowie Vor- und Nachteile erläutert. Ausführlicher wird dabei ein zweistufiger Matching-Algorithmus vorgestellt, der sich in den vergangenen Jahren als Standardverfahren etabliert hat. Im IWH ist diese Methode weiterentwickelt worden, um die Aussagefähigkeit der Evaluationsergebnisse zu verbessern.

Dieses weiterentwickelte Verfahren wurde der Analyse der Effekte von Arbeitsbeschaffungs- und Strukturanpassungsmaßnahmen, die im vorangehenden Beitrag beschrieben sind, zugrundegelegt.[22]

Die Evaluation arbeitsmarktpolitischer Maßnahmen hat in den letzten Jahren wesentlich an Bedeutung gewonnen. Die in Evaluationsstudien erzielten Ergebnisse sind dem interessierten Leser häufig geläufig. Viel weniger bekannt ist hingegen die methodische Herangehensweise.

Die folgenden Ausführungen geben einen Überblick über die Grundidee der Evaluation sowie ausgewählte Methoden ihrer empirischen Umsetzung.

Das Grundproblem der Evaluation arbeitsmarktpolitischer Maßnahmen

Das Ziel mikroökonomischer Evaluationsstudien ist die Feststellung der Auswirkungen arbeitsmarktpolitischer Maßnahmen auf die (Wieder-)Beschäftigungschancen der teilnehmenden Personen.

[21] Eine Übersicht über Ausgaben für die einzelnen arbeitsmarktpolitischen Instrumente findet sich unter http://www1.arbeitsamt.de/hst/services/statistik/200212/iiia4/ampb.pdf.

[22] Vgl. auch REINOWSKI, E.; SCHULTZ, B.; WIEMERS, J.: Verschlechterung der Beschäftigungschancen durch Teilnahme an Arbeitsbeschaffungs- und Strukturanpassungsmaßnahmen, in diesem Heft, S. 184-190.

Um individuelle Effekte ermitteln zu können, müsste nach Abschluss einer Maßnahme die Arbeitsmarktsituation eines Teilnehmers mit der hypothetischen Situation im Falle einer Nichtteilnahme verglichen werden. Beide Situationen sind für ein und dieselbe Person allerdings nicht beobachtbar. Für die beobachtete Beschäftigungssituation eines Teilnehmers fehlt also der Vergleichsmaßstab.

Um dieses Problem zu lösen, kann anstelle des individuellen Vergleichs ein Durchschnittseffekt für die gesamte Teilnehmergruppe ermittelt werden. Das größte Problem dabei stellt die Konstruktion einer geeigneten Vergleichsgröße dar.

Hierzu kann beispielsweise die Beschäftigungssituation eines Teilnehmers zu einem früheren Zeitpunkt oder die eines Nichtteilnehmers verwendet werden. Allerdings muss dann in geeigneter Weise berücksichtigt werden, dass sich Teilnehmer und Nichtteilnehmer nicht nur dadurch unterscheiden, dass die eine Gruppe an einer Maßnahme teilgenommen hat und die andere nicht. Wenn die Unterschiede in anderen beschäftigungsrelevanten Merkmalen nicht in die Betrachtung einbezogen werden, wird das Ergebnis durch die sog. Selektionsverzerrung verfälscht.

Diese beschäftigungsrelevanten Unterschiede lassen sich in zwei Gruppen einteilen. So müssen beobachtbare Merkmale wie z. B. Alter, Geschlecht oder Ausbildung und nicht erfasste oder nicht beobachtbare Faktoren wie Motivation bei einer Evaluation berücksichtigt werden.

Überblick über Methoden zur Lösung des Selektionsproblems

Zur Berücksichtigung der Selektionsverzerrung wurden verschiedene Verfahren entwickelt, die sich im Wesentlichen darin unterscheiden, in welcher Weise die Verzerrung berücksichtigt wird. Eine Möglichkeit ist die Beseitigung der Verzerrung bei der Konstruktion der Vergleichsgröße selbst. Dabei sind weder Verteilungsannahmen noch Annahmen über funktionale Zusammenhänge für eine Schätzung des Maßnahmeeffekts nötig.

Die Vergleichsgröße kann mit Hilfe verschiedener Verfahren gebildet werden.

Bei der intuitiv am einfachsten zugänglichen Methode – dem *Vorher-Nachher-Vergleich* – wird eine Person mit sich selbst verglichen. Zur Ermittlung des Maßnahmeeffekts wird die Situation der Teilnehmer vor der Maßnahme mit derjenigen danach gegenübergestellt. Grundlegend dafür ist die Annahme, dass die Beschäftigungswahrscheinlichkeit zwischen den betrachteten Zeitpunkten nur durch die Teilnahme an einer Maßnahme und nicht durch andere Einflüsse wie konjunkturelle Schwankungen oder zeitbedingte Veränderungen beeinflusst wird.

Ein großer Vorteil des Vorher-Nachher-Vergleichs besteht in dem relativ geringen Datenbedarf. Es sind keine Informationen über Nichtteilnehmer erforderlich, sondern lediglich Daten der Teilnehmer für jeweils einen Zeitpunkt vor und nach der Maßnahme. Gegenüber einer Verzerrung durch beobachtbare und im Untersuchungszeitraum konstante nicht beobachtbare Merkmale ist dieses Verfahren robust. Allerdings ist er sehr anfällig gegenüber der Verletzung der grundlegenden Annahme. Ein Beispiel dafür ist der sog. Ashenfelter Dip, die Beobachtung, dass die Beschäftigungswahrscheinlichkeit der Teilnehmer kurz vor Maßnahmebeginn sinkt.[23] Wenn diese Veränderung nur vorübergehend und nur bei Teilnehmern zu beobachten ist, wird der durchschnittliche Maßnahmeeffekt überschätzt.

Simulationsstudien haben gezeigt, dass der Vorher-Nachher-Vergleich deutlich verzerrte Schätzergebnisse liefert, sobald gesamtwirtschaftliche Veränderungen, lebenszyklusbedingte Änderungen, Ashenfelter Dip oder eine Kombination aus mehreren „Störfaktoren“ auftritt.[24]

Eine andere Möglichkeit der Ermittlung des Maßnahmeeffekts ist der Vergleich der Teilnehmergruppe mit einer geeigneten Gruppe von Nichtteilnehmern.

Der einfachste Vertreter dieses Vergleichs zweier unterschiedlicher Personengruppen ist der *Kreuz-Vergleich*. Hier wird die Beschäftigungssituation von Teilnehmern und Nichtteilnehmern nach Be-

---

[23] Sehr anschaulich wird der Ashenfelter Dip und seine Auswirkungen auf Schätzergebnisse in HUJER, R.; CALIENDO, M.; RADIC, D.: Nobody Knows... How do Different Evaluation Estimators Perform in a Simulated Labour Market Experiment?, Diskussionspapier der Universität Frankfurt/Main 2001, S. 7 f., erläutert.

[24] Zu diesem Ergebnis kommen HECKMAN, J. J.; SMITH, J.: Pre-Program Earnings Dip and the Determinant of Participation in a Social Program: Implications for Simple Program Evaluation Strategies, in: Economic Journal, Vol. 109, No. 457, 1999, S. 322 in Verbindung mit Tabelle 1.

endigung der Maßnahme verglichen. Dazu wird angenommen, dass im Gruppendurchschnitt die Beschäftigungswahrscheinlichkeit der Nichtteilnehmer mit der hypothetischen der Teilnehmer übereinstimmt.

Die Anwendung des Kreuz-Vergleichs erfordert den Zugang zu Querschnittsdaten für Teilnehmer- und Nichtteilnehmergruppe.

Im Gegensatz zum Vorher-Nachher-Vergleich ist er robust gegenüber gesamtwirtschaftlichen Veränderungen und lebenszyklusbedingten Beschäftigungsänderungen, wenn diese die Teilnehmer und die Kontrollgruppenmitglieder gleichermaßen betreffen. Auch der Ashenfelter Dip stellt kein Problem dar, da nur Daten nach Maßnahmeende betrachtet werden.

Allerdings dürfen keine unbeobachtbaren Unterschiede zwischen den beiden Gruppen bestehen. Wenn beispielsweise die Motivation einiger Personen Einfluss auf die Beschäftigungswahrscheinlichkeit und die Maßnahmeteilnahme hat, ist der Maßnahmeeffekt-Schätzer verzerrt.

Ein Schätzer, der in der Lage ist, auch unbeobachtbare Merkmale zu berücksichtigen, ist die *Differenz-von-Differenzen-Methode*. Sie ist eine Kombination aus Vorher-Nacher-Vergleich und Kreuz-Vergleich. Beim Vergleich der Teilnehmergruppe mit geeigneten Nichtteilnehmern werden die Veränderungen beider Gruppen innerhalb des Untersuchungszeitraums betrachtet.

Dazu muss im Gruppendurchschnitt die Differenz der Beschäftigungsniveaus der Nichtteilnehmer für jeweils einen Zeitpunkt vor und nach der Maßnahme der hypothetischen Beschäftigungsdifferenz der Teilnehmer entsprechen, so die Zentralannahme.

Dieser Schätzer erfordert Zugang zu umfangreichen Längsschnittdaten für Teilnehmer- und Nichtteilnehmergruppe. Ein Vorteil des Verfahrens liegt in seiner Robustheit gegenüber gesamtwirtschaftlichen Veränderungen und lebenszyklusbedingten Änderungen der Beschäftigungswahrscheinlichkeit, wenn diese die Teilnehmer und die Nichtteilnehmer gleichermaßen betreffen.[25] Außerdem ist er in der Lage, Selektionsverzerrung wegen unbeobachtbarer Heterogenitäten zu beseitigen, wenn sie durch zeitinvariante Charakteristika hervorgerufen wird.

Allerdings kann der Maßnahmeeffekt nur unverzerrt ermittelt werden, wenn kein Ashenfelter Dip auftritt. Die Ergebnisse sind also sehr sensitiv gegenüber der Wahl der Zeitpunkte, zu denen das Vor- bzw. Nach-Maßnahme-Beschäftigungsniveau beobachtet wird.[26]

Das in letzter Zeit am häufigsten verwendete Verfahren ist die *Matching-Methode*. Auch sie vergleicht die Teilnehmergruppe mit geeigneten Nichtteilnehmern. Im Unterschied zum Kreuz-Vergleich werden allerdings für jeden einzelnen Teilnehmer geeignete Nichtteilnehmer ermittelt. Die wichtigste Annahme dabei ist, dass die ausgewählten Nichtteilnehmer die gleichen Beschäftigungsaussichten wie der entsprechende Teilnehmer hätten, wenn sie auch an einer Maßnahme teilnehmen würden.

Diese Annahme kann mit verschiedenen Matchingprozessen erfüllt werden. Sie unterscheiden sich im Wesentlichen in der Wahl des Kriteriums, mit dessen Hilfe die Ähnlichkeit zwischen Teilnehmern und Nichtteilnehmern festgestellt wird und der Anzahl der zugeordneten Nichtteilnehmer.[27]

Das IWH verwendet ein zweistufiges Verfahren, das jedem Teilnehmer genau einen Nichtteilnehmer zuordnet, der als „statistischer Zwilling" des Teilnehmers in die Kontrollgruppe aufgenommen wird.

Der Matchingprozess

In Vorbereitung des eigentlichen Matching wird ein Katalog aller Merkmale zusammengestellt, die für die Bestimmung der Ähnlichkeit von Personen genutzt werden sollen.

Im ersten Schritt wird mit Hilfe der ausgewählten Merkmale der sog. Propensity Score[28] geschätzt. Anhand dieses eindimensionalen Indikators wird für jeden Teilnehmer eine Untergruppe

[25] Wenn die Einkommensänderungen in beiden Gruppen unterschiedlich ausfallen, liefert die Differenz-von-Differenzen-Methode verzerrte Schätzergebnisse. Das wurde bei einem Vergleich der Ergebnisse dieser Methode mit denen eines sozialen Experiments nachgewiesen. Siehe dazu HECKMAN, J. J.; SMITH, J., a. a. O., S. 315.

[26] Siehe dazu beispielsweise HECKMAN, J. J.; SMITH, J., a. a. O., S. 325.

[27] Eine übersichtliche Einführung in die verschiedenen Matching-Verfahren findet sich in HUJER, R.; CALIENDO, M.; RADIC, D., a. a. O., S. 9 ff.

[28] Der Propensity Score beschreibt die Wahrscheinlichkeit, eine Maßnahmeteilnahme zu beobachten.

Abbildung:
Ablaufschema des iterativen Verfahrens zur Zuordnung möglichst ähnlicher Nichtteilnehmer zu den untersuchten Teilnehmern

Quelle: IWH

mit den ihm ähnlichsten Nichtteilnehmern gebildet.[29]

Zu Beginn des zweiten Schritts wird für die Wahl der Vergleichsperson aus den gebildeten Untergruppen ein mehrdimensionales Ähnlichkeitskriterium festgelegt, die sog. Mahalanobisdistanz.[30] Für die Zuordnung werden zwei verschiedene Techniken angewendet und miteinander verglichen: die Standardtechnik und ein vom IWH entwickeltes iteratives Verfahren. Beide Algorithmen ordnen jedem Teilnehmer genau einen Nichtteilnehmer zu. Ein Nichtteilnehmer kann dabei nicht als Vergleichsperson für mehrere Teilnehmer eingesetzt werden. So wird sichergestellt, dass Teilnehmer- und Kontrollgruppe die gleiche Größe haben.

Bei der Standard-Zuordnung wird nach dem Zufallsprinzip eine Reihenfolge festgelegt, nach der jeder Teilnehmer dem ihm ähnlichsten Nichtteilnehmer aus dem Pool der betrachteten Personen zugeordnet wird. Dabei kann jedoch nicht die Zuordnung der ähnlichsten Vergleichsperson für jeden Teilnehmer garantiert werden, da die gebildeten Zweiergruppen aus diesem Pool entfernt werden. Für die zuletzt zuzuordnenden Teilnehmer können die ähnlichsten Nichtteilnehmer dann schon vergeben sein. Bei einer ungünstigen Datenlage besteht sogar die Gefahr, dass für die zuletzt zuzuordnenden Teilnehmer keine Vergleichsperson mehr gefunden werden kann, weil die ihnen ähnlichen Nichtteilnehmer schon anderen Teilnehmern zugeordnet worden sind. Diese Teilnehmer können dann in der weiteren Untersuchung nicht mehr berücksichtigt werden.

Um die suboptimale Zuordnung der Vergleichspersonen und den Verlust von Beobachtungen zu vermeiden, ist ein iterativer Prozess entwickelt worden, der den Austausch bereits zugeordneter Personen ermöglicht (vgl. dazu die Abbildung). Den Ausgangspunkt bilden Zweiergruppen aus jeweils einem Teilnehmer und einem zufällig aus seiner Untergruppe ausgewählten Nichtteilnehmer. Alternativ kann auch das Ergebnis der Standard-Zuordnung als Grundlage genutzt werden. Für alle diese Zweiergruppen wird die Summe der quadrierten Mahalanobisdistanzen ermittelt. Die Minimierung dieser Summe ist Ziel des Prozesses, der nach folgenden Regeln abläuft: Es werden für einen zufällig gewählten Teilnehmer alle weiteren möglichen Vergleichspersonen festgestellt und davon eine zufällig ausgewählt. Im einfacheren Fall ist der Nichtteilnehmer noch keinem anderen Teilnehmer zugewiesen. Er wird gegen den ursprünglich zugeordneten Nichtteilnehmer ausgetauscht, wenn sich dadurch die Summe der quadrierten Mahalanobisdistanzen verringert. Ist dagegen der ausgewählte Nichtteilnehmer bereits einem anderen Teilnehmer zugeordnet, wird zusätzlich überprüft, ob dieser Teilnehmer noch weitere mögliche Vergleichspersonen hat. Wenn weitere Personen vorhanden sind, wird nach dem oben beschriebenen Muster ausgetauscht. Dieser Prozess wird so oft wiederholt, bis mit einer vorher festgelegten Anzahl von Durchläufen keine Verringerung der quadrierten Distanzsumme mehr erreicht werden kann.

Der Vergleich der Ergebnisse beider Verfahrens zeigt eine Erhöhung der Ähnlichkeit zwischen Teilnehmern und jeweils zugeordneten Nichtteilnehmern im Vergleich zum Standardverfahren. Damit kann die Qualität der Evaluation arbeitsmarktpolitischer Maßnahmen erhöht werden, da mit der ermittelten Vergleichsgröße die hypothetische Situation der Nichtteilnahme für die Teilnehmer besser abgebildet wird.

Eva.Reinowski@iwh-halle.de
Birgit.Schultz@iwh-halle.de
Jürgen.Wiemers@iwh-halle.de

[29] Der Einsatz eines eindimensionalen Indikators, eines sog. Balancing Scores, zur Feststellung der Ähnlichkeit von zwei Personen vermeidet das sog. Dimensionsproblem: Mit jedem zusätzlich betrachteten Merkmal erhöht sich die Qualität des Matchingergebnisses, aber die Anzahl der zu überprüfenden möglichen Übereinstimmungen zwischen Teilnehmer und potentiellen Vergleichspersonen steigt exponentiell. Vgl. HUJER, R.; CALIENDO, M.; RADIC, D., a. a. O., S. 10. Hier findet sich ein anschauliches Beispiel für dieses Problem.

[30] Die Mahalanobisdistanz ist ein quantitatives Maß für die Unterschiede zweier Personen in den einzelnen betrachteten Merkmalen.