Institut für Arbeitsmarkt- und Berufsforschung

Die Forschungseinrichtung der Bundesagentur für Arbeit

IAB

IAB-Discussion Paper 7/2009

Beiträge zum wissenschaftlichen Dialog aus dem Institut für Arbeitsmarkt- und Berufsforschung

Fehlende Daten beim Record Linkage von Prozess- und Befragungsdaten

Ein empirischer Vergleich ausgewählter Missing Data Techniken

Gerhard Krug

Fehlende Daten beim Record Linkage von Prozess- und Befragungsdaten

Ein empirischer Vergleich ausgewählter Missing Data Techniken

Gerhard Krug (IAB)

Mit der Reihe „IAB-Discussion Paper“ will das Forschungsinstitut der Bundesagentur für Arbeit den Dialog mit der externen Wissenschaft intensivieren. Durch die rasche Verbreitung von Forschungsergebnissen über das Internet soll noch vor Drucklegung Kritik angeregt und Qualität gesichert werden.

The “IAB-Discussion Paper” is published by the research institute of the German Federal Employment Agency in order to intensify the dialogue with the scientific community. The prompt publication of the latest research results via the internet intends to stimulate criticism and to ensure research quality at an early stage before printing.

Inhaltsverzeichnis

Zusammenfassung

Zum Vergleich ausgewählter Missing Data Techniken nutzt dieses Papier eine Befragung, in der u. a. die Zustimmung zum Record Linkage der Befragungs- mit administrativen Prozessdaten abgefragt wurde. Bei nicht zustimmenden Befragten, werden ihre gegebenen Antworten auf „fehlend“ gesetzt, um so pseudo-fehlende Werte auf Basis eines empirischen (im Vergleich zu einem statistisch simulierten) Ausfallmechanismus zu erzeugen. Eine OLS Regression wird durchgeführt und dem Datenausfall wird jeweils durch eine Complete Case Analyse (CCA), Multiple Imputation (MI) und zwei Varianten des Heckmans Sample Selection Models (SSM) begegnet. Die Ergebnisse werden mit einer Regression auf Basis der vollständigen Daten verglichen, welche die „wahren“ Regressionsergebnisse liefert (Benchmark). Alle Verfahren führen zu nur wenigen Abweichungen vom Benchmark. Wenn nur eine unabhängige Variable fehlende Werte aufweist, liegt die MI näher zum Benchmark, wenn die abhängige ausfallbelastet ist, die CCA gefolgt von der Two-Step Variante des SSM. Bei fehlenden Werten in vielen oder allen unabhängigen Variablen zeigen sich alle Verfahren ähnlich geeignet zur Korrektur der Ausfälle, mit Ausnahme der Maximum Likelihood Variante des SSM.

Abstract

To compare different missing data techniques, in this paper I use a survey where participants were among other things asked permission for combining the survey with administrative data (record linkage). For those who refuse their permission I set their survey answers to missing, creating pseudo-missing data due to an empirical relevant but unknown mechanism (compared to the statistical simulation of a missing data process). OLS Regression is performed using Complete Case Analysis (CCA), Multiple Imputation (MI) and two versions of Heckman’s Sample Selection Model (SSM) to correct for the pseudo-missing data. Their results are compared to a regression based on the complete data set (Benchmark), that gives us the “true” regression parameters. Results: All missing data techniques under analysis show only small deviations from the benchmark. If only one independent variable contains missing values, MI performs best. If the dependent variable has missing information, CCA and the Two-Step SSM perform better than MI. If missing data is a problem in many or all independent variables, all techniques except for the Maximum likelihood SSM perform equally well.

JEL Klassifikation: C81, J20

Keywords: Missing Data; Multiple Imputation; Sample Selection Model; Record Linkage

Für wertvolle Hinweise danke ich Jörg Drechsler, Hans Kiesl, Martin Spiess und den Teilnehmern der Session "Combining Data from Different Sources" auf der ISA-RC33 7th International Conference in Neapel. Verbliebene Fehler liegen in meiner Verantwortung.

1 Einleitung

Standardisierte Befragungen stellen ein zentrales Element der empirischen Sozialforschung dar. Durch die Entstehung von Forschungsdatenzentren und die Aufbereitung administrativer Daten zu Scientific Use Files rücken aber auch sogenannte prozessproduzierte Daten in den Blick der Forschung (Wirth/Müller 2004; Allmendinger/Kohlmann 2005). Angesichts der Tatsache, dass die Vorteile beider Datenquellen zum Teil auf unterschiedlichen Gebieten liegen (vgl. Hartmann/Krug 2009), bietet es sich an, durch ihre Kombination die Aussagekraft empirischer Analysen zu erweitern. Eine Möglichkeit besteht in der Datenverknüpfung (Record Linkage), bei der auf der individuellen Ebene Befragungsdaten mit prozessproduzierten Informationen zum selben Individuum angereichert werden.[1] In vielen Fällen gilt allerdings, dass vor einer solchen Verknüpfung die Erlaubnis der betroffenen Personen einzuholen ist. Obwohl erfahrungsgemäß die Zustimmungsbereitschaft der Befragten relativ hoch ist (z. B. Hartmann et al. 2008: 57), wird diese natürlich auch von einem Teil der Befragten verweigert. Diese Personen weisen dann in dem angereicherten Datensatz bei den entsprechenden Variablen fehlende Werte auf.

Für die empirische Forschung mit solchen Daten stellt sich damit die Frage des Umgangs mit den fehlenden Werten. Die einfachste Lösung besteht darin, für Analysen nur die vollständigen Fälle zu verwenden. Dies setzt jedoch einen zufälligen Ausfall der nicht verwendeten Beobachtungen voraus. Ist dies nicht der Fall, werden Schätzungen etwa von Mittelwerten oder Regressionskoeffizienten verzerrt sein. Statistische Verfahren wie Multiple Imputation oder Sample Selection Models versprechen hier Abhilfe. Dabei gehen sie von bestimmten Annahmen über den Datenausfallprozess aus, so dass bei Nichterfüllung dieser Annahmen das Ziel unverzerrter Schätzungen aber eventuell verfehlt wird.

Da diese Annahmen im konkreten Anwendungsfall meist nicht testbar sind, ist die Entscheidung für das eine oder andere Verfahren zum Umgang mit fehlenden Werten (sog. Missing Data Techniken) oft schwierig. Die vorliegende Arbeit prüft im Rahmen einer Fallstudie einige ausgewählte Verfahren, und zeigt auf, inwiefern sie bei der Korrektur von Stichprobenausfällen zu unterschiedlichen Ergebnissen führen. Aufgrund des gewählten Analysedesigns kann an einem ausgewählten Fall nicht nur nachvollzogen werden, ob die Verfahren in der Forschungspraxis zum selben, sondern auch zum *richtigen* Ergebnis gelangen. Es wird hierzu eine Befragung genutzt, bei der die Zustimmung zur Verknüpfung mit administrativen Daten abgefragt wurde. Solche empirischen Vergleiche sind zum Beispiel geeignet, um die Robustheit empirischer Forschungsergebnisse im Hinblick auf das gewählte Verfahren unter realistischen Anwendungsbedingungen zu analysieren (vgl. etwa Ridder 1992). Im Unterschied zu vollständig simulierten Daten ermöglicht der echte Daten-

[1] Vom Record Linkage ist das statistische Matching (Rässler 2002) zu unterscheiden, bei dem Informationen von aus statistischer Sicht möglichst ähnlichen Individuen miteinander verknüpft werden.

satz in Kombination mit dem gewählten Analysedesign, das Fehlen von Prozessdaten aufgrund tatsächlicher empirischer Teilnahmeentscheidungen von Befragten nachzuahmen. So liefern die vorliegenden Analysen im Gegensatz zu Simulationen einen Hinweis bezüglich der Eignung der Missing Data Verfahren, den *empirisch* aufgrund bestehender Datenschutzregelungen auftretenden Datenausfall beim Record Linkage von Prozess- und Befragungsdaten auszugleichen.

Die vorliegende Arbeit ist hierzu wie folgt aufgebaut. Zunächst werden im folgenden Abschnitt Annahmen zu verschiedenen Ausfallmechanismen und ihnen korrespondierende Verfahren zum Umgang mit fehlenden Werten vorgestellt. Im Anschluss wird der empirische Vergleich durchgeführt, wobei zunächst die Datenbasis und das Design des Vergleichs vorgestellt, die Implementation der Missing Data Techniken besprochen und schließlich die Ergebnisse präsentiert werden. In Abschnitt erfolgen abschließende Schlussfolgerungen aus dem empirischen Vergleich der Verfahren.

2 Ausfallmechanismen und Missing Data Techniken

Die Zusammenspielung von Prozess- und Befragungsdaten kann unterschiedlichen Zwecken dienen. Im Folgenden wird davon ausgegangen, dass mit den verknüpften Prozess- und Befragungsdaten Regressionsanalysen durchgeführt werden sollen, z. B. um den Erfolg arbeitsmarktpolitischer Maßnahmen im Hinblick auf bestimmte Erfolgsindikatoren zu untersuchen (vgl. zum Folgenden Horton/Lipsitz 2001; Horton/ Kleinman 2007). **Y** bezeichnet den Vektor der abhängigen Variablen und **X** den Vektor der unabhängigen Variablen. Für eine Beobachtungseinheit sind die Werte dieser Variablen entweder beobachtet oder unbeobachtet. $\mathbf{Y}^{obs}$ ist die beobachtete Komponente der Zielvariablen und $\mathbf{X}^{obs}$ die der Regressoren. Entsprechend sind $\mathbf{Y}^{mis}$ und $\mathbf{X}^{mis}$ die unbeobachteten Komponenten. Für spätere Zwecke werden abhängige und unabhängige Variablen zusammen mit **Z** bezeichnet, so dass $\mathbf{Z}^{mis}$ = ($\mathbf{Y}^{mis}$, $\mathbf{X}^{mis}$) und $\mathbf{Z}^{obs}$ = ($\mathbf{Y}^{obs}$, $\mathbf{X}^{obs}$). Das inhaltliche Interesse der Analyse mit verknüpften Daten bezieht sich auf die Regressionsparameter β, welche die bedingte Verteilung von **Y** gegeben **X** bestimmen: $f(\mathbf{Y} \mid \mathbf{X}, \beta)$.

Grundsätzlich lassen sich hinsichtlich des Datenausfalls im Allgemeinen und so auch im Fall des Record Linkage drei unterschiedliche Situationen unterscheiden (Rubin 1987; Little/Rubin 1987; Collins/Schafer/Kam 2001): missing at random (MAR), missing completely at random (MCAR) und missing not at random (MNAR). Diese Situationen unterscheiden sich danach, welche Annahmen über die Beziehung zwischen den in der konkreten inhaltlichen Analyse relevanten Variablen und den Determinanten des Ausfallprozesses gerechtfertigt sind. Sei **R** ein Indikator dafür, ob ein Element von **Z** beobachtet oder unbeobachtet ist, mit R_j = 1 falls das j-te Element von **Z** beobachtet ist und R_j = 0 sonst und sei ϕ der Parametervektor, der den Ausfallprozess kennzeichnet. Missing completely at random (MCAR) ist demnach definiert als

$$P(\mathbf{R} \mid \mathbf{Z}) = P(\mathbf{R} \mid \mathbf{Z}^{obs}, \mathbf{Z}^{mis}) = P(\mathbf{R} \mid \phi) \quad (1)$$

wobei ϕ und β als distinkt angenommen werden. Intuitiv bedeutet MCAR: „the process generating missing values bears no statistical relationship (e.g. correlations) with our variables of interest“ (Collins/Schafer/Kam 2001: 333). Diese Annahme erscheint jedoch gerade dann, wenn der Ausfall auf Entscheidungen der befragten Individuen beruht, als problematisch. Die missing at random (MAR) Annahme ist dagegen weniger restriktiv und lautet:

$$P(\mathbf{R} \mid \mathbf{Z}) = P(\mathbf{R} \mid \mathbf{Z}^{obs}, \phi) \tag{2}$$

Die Intuition hinter der MAR ist, dass der Ausfallprozess zwar mit den interessierenden Variablen zusammenhängt, diese Beziehung aber vollständig von den beobachteten Daten in der Analyse erfasst wird. Graham (2009: 553) spricht daher auch von „conditionally missing at random“. Im Fall MCAR wie auch MAR wird der Ausfallprozess (missing data mechanism) als ignorierbar (ignorable) bezeichnet. Schließlich besagt die MNAR Annahme, dass der Ausfallprozess $P(\mathbf{R} \mid \mathbf{Z})$ nicht weiter vereinfacht werden kann, da er auch auf unbeobachteten Daten beruht; er ist nicht ignorierbar (nonignorable) (Little, Rubin 1987).

Es gibt eine Reihe verschiedener Möglichkeiten, mit fehlenden Daten umzugehen. Diese lassen sich danach unterscheiden, welche der drei genannten Annahmen zum Datenausfall mindestens erfüllt sein muss, damit sie anwendbar sind.

Complete Case Analyse (CCA): Missing completely at random

Die Complete Case Analyse ist zunächst die unkomplizierteste Möglichkeit, mit fehlenden Daten umzugehen. Hier werden für die Analyse nur diejenigen Fälle verwendet, für die alle Variablen beobachtete Werte aufweisen. Im vorliegenden Fall wären das also ausschließlich diejenigen Personen, welche dem Record Linkage zustimmen. Dabei wird allerdings davon ausgegangen, dass die Teilstichprobe der Zustimmer eine einfache Zufallsstichprobe aus allen Befragten darstellt und damit der Datenausfall missing completely at random ist. Ist diese Annahme tatsächlich erfüllt, können auf Basis der verfügbaren Fälle unverzerrte Schätzungen vorgenommen werden, wenn auch die Schätzung wegen geringerer Fallzahl an Effizienz verliert. Ist dies nicht der Fall, führt etwa eine Regressionsanalyse zu verzerrten Parameterschätzungen.

Multiple Imputation (MI): Missing at random

Unter Multiple Imputation (Rubin 1976, 1987; Weins 2006) versteht man ein Verfahren, bei dem die fehlenden Werte in den Daten mit m>1 plausiblen Werten ersetzt werden, wodurch ebenso viele Datensätze entstehen. Das Verfahren setzt voraus, dass die Beziehung zwischen Datenausfall und ausfallbehafteter Variable vollständig von beobachteten Daten abhängt (MAR). Das Vorgehen bei der multiplen Imputation kann in drei Teilschritte zerlegt werden: Imputation, Datenanalyse und Kombination der Ergebnisse.

Im *Imputationsschritt* werden zunächst mit m>1 mehrere plausible Werte für die fehlenden Werte erzeugt. Die MAR-Annahme garantiert dabei, dass $\left(\mathbf{Z}^{\{1\}}, \mathbf{Z}^{\{2\}}, ..., \mathbf{Z}^{\{m\}}\right)$ ergänzte Datensätze aus der Verteilung $f(\mathbf{Z}^{mis} \mid \mathbf{Z}^{obs})$ erzeugt werden können, da nach der Konditionierung auf $\mathbf{Z}^{obs}$ der Datenausfall - im Bezug auf die betrachteten Variablen - zufällig erfolgt. Es existiert eine Vielzahl von Varianten zur Erzeugung der Imputationen, von Propensity Score Methoden über predictive mean matching, Diskriminanzanalysen bis hin zu logistischen Regressionen (vgl. Horton/Kleinman 2007). Bei komplexeren Ausfallmustern bieten sich meist Markov Chain Monte Carlo (MCMC) Methoden an. Hier wird eine Markov-Kette[2] erzeugt, um Ziehungen aus der Posteriorverteilung $f(\mathbf{Z}^{mis} \mid \mathbf{Z}^{obs})$ zu simulieren. Die Implementation der MCMC Methode kann über den IP-Algorithmus erfolgen (Schafer 1997), der zwischen Imputations- und Parameterschritt iteriert. Basierend auf der Parameterschätzung $\phi^{(t)}$ in der t-ten Iteration, wird im Imputationsschritt $\mathbf{Z}^{mis,(t+1)}$ aus $f(\mathbf{Z} \mid \mathbf{Z}^{obs}, \phi^{(t)})$ gezogen und im anschließenden Parameterschritt wird $\phi^{(t+1)}$ aus $f(\mathbf{Z} \mid \mathbf{Z}^{obs}, \mathbf{Z}^{mis,(t+1)})$ gezogen, usw. Dies erzeugt eine Markov-Kette $\left(\left\{\mathbf{Z}^{(1)}, \phi^{(1)}\right\}, \left\{\mathbf{Z}^{(2)}, \phi^{(2)}\right\}, ..., \left\{\mathbf{Z}^{(t+1)}, \phi^{(t+1)}\right\}, ...\right)$ welche schließlich zur gesuchten Posteriorverteilung konvergiert (Schafer 1997: 72).

Unabhängig davon wie die Imputationen konkret erzeugt werden, erfolgt im nächsten Schritt die *Datenanalyse* stets in den m generierten ergänzten Datensätzen unter Verwendung von Standardverfahren der statistischen Analyse, z. B. also Regressionsanalysen. Im *Kombinationsschritt* werden dann die Ergebnisse (die Schätzungen für die Regressionsparameter β_m) der m separaten Analysen aus den verschiedenen imputierten Datensätzen gemäß einfacher Kombinationsregeln (Rubin 1987) miteinander verknüpft:

Es sei $m = 1,...M$ die Zahl der Imputationen, dann ist der MI-Schätzer für die Regressionskoeffizienten der einfache Durchschnitt über alle M Imputationen: $\overline{\beta}_M = \frac{1}{M}\sum_{m=1}^{M} \hat{\beta}_m$. Zur Bestimmung der Standardfehler wird zunächst die within-imputation variance $\overline{W}_M = \frac{1}{M}\sum_{m=1}^{M} W_m$ berechnet, mit $W_m = V\hat{a}r(\hat{\beta}_m)$ in der m-ten Imputation, sowie die between-imputation variance $B_M = \frac{1}{M-1}\sum_{m=1}^{M}\left(\hat{\beta}_m - \overline{\beta}_M\right)^2$. Der MI-Schätzer für die Gesamtvarianz kombiniert beide Werte auf folgende Weise: $V_M = \overline{W}_M + \frac{M+1}{M} B_M$.

---

[2] Eine Markov-Kette ist eine Sequenz von Zufallsvariablen, in der die Verteilung eines jeden Elementes vom Wert des vorherigen abhängt. Bei der MCMC Methode wird eine Kette erzeugt, die lange genug ist, damit die Elemente zu einer stabilen (stationären) Verteilung konvergieren.

Sample Selection Model (SSM): Missing not at random

Kann man nicht davon ausgehen, dass alle relevanten Einflüsse auf die Zustimmung in den beobachteten Daten erfasst sind, ist die Schätzung eines Sample Selection Models (zum Folgenden Heckman 1979; Engelhardt 1999) eine mögliche Alternative zur Multiplen Imputation. Meist wird das SSM eingesetzt um den Ausfall in einer einzigen Variable zu korrigieren, typischer Weise der Abhängigen Y_1, während die Kontrollvariablen vollständig beobachtet werden. Das Verfahren ist aber grundsätzlich auch bei Ausfällen in mehreren Variablen anwendbar.

Aus der MNAR Annahme ergibt sich, dass eine OLS-Schätzung der Regressionsgleichung

$$Y_{1i} = X_i\beta + \varepsilon_i \tag{3}$$

zu verzerrten Parameterschätzungen führt. Um dies zu vermeiden, wird eine zweite Gleichung formuliert, die den Selektionsprozess beschreibt, soweit dieser durch die vorhandenen Daten abzubilden ist.

$$D_i^* = C_i\alpha + \nu_i \tag{4}$$

Dabei ist D* eine latente Variable, etwa die latente Bereitschaft, dem Record Linkage zuzustimmen, und *i* ein Personenindex. Übersteigt die latente Variable einen bestimmten Wert (z. B. 0), dann stimmt Person *i* dem Zusammenspielen zu und sonst nicht:

$$D_i = \begin{cases} 1 \text{ if } D_i^* > 0, \\ 0 \text{ sonst} \end{cases}$$

Demnach wird Y_1 nur für solche Personen beobachtet, für die $\nu_i > -C_i\alpha$ ist, weshalb der Erwartungswert von Y_1 in der Teilpopulation nicht $X_i\beta$ ist, sondern $E(Y_{1i} \mid D_i = 1, X_i) = X_i\beta + E(\varepsilon_i \mid D_i = 1_i) = X_i\beta + E(\varepsilon_i \mid \nu_i > -C_i\alpha)$.

Damit stellt sich die Schätzung des Erwartungswertes von Y_1 für die gesamte Population als ein Problem des Fehlens einer Variable dar. Unter der Annahme, dass die Störgrößen in den Gleichungen (3) und (4) einerseits bivariat normalverteilt sind $(\varepsilon_i, \nu_i) \sim N(0,0,\sigma_\varepsilon^2, \sigma_\nu^2, \rho_{\varepsilon\nu})$ und andererseits unabhängig von X und C, gilt $E(\varepsilon_i \mid \nu_i > -C_i\alpha) = \rho_{\varepsilon\nu}\sigma_\varepsilon \frac{\phi(C_i\alpha)}{\Phi(C_i\alpha)}$ (vgl. Heckman 1979). Die Variable $\frac{\phi(C_i\alpha)}{\Phi(C_i\alpha)}$ wird als inverse Mills Ratio oder non selection hazard bezeichnet.

Im so genannten Two-Step Verfahren, im Folgenden auch SSM(2) wird sie in einem ersten Schritt aus der vorhandenen Stichprobe mit $\frac{\phi(C_i\hat{\alpha})}{\Phi(C_i\hat{\alpha})}$ geschätzt werden, indem eine Probitregression der Selektionsgleichung (4) durchgeführt wird.

Im zweiten Schritt wird sie in die Regressionsgleichung für Y_1 eingesetzt. Eine OLS-Schätzung der resultierenden Regressionsgleichung

$Y_{1i} = X_i\beta + \rho_{\varepsilon v}\sigma_\varepsilon \frac{\phi(C_i\hat{\alpha})}{\Phi(C_i\hat{\alpha})} + \varepsilon_i$ liefert dann eine unverzerrte Schätzung der Regressionsparameter β.

Um Kollinearitätsprobleme zu vermeiden sollte dabei C mindestens ein Element enthalten, das nicht auch in X enthalten ist (Instrumentvariable oder exclusion restriction, vgl. Puhani 2000). Anstatt des Two-Step - Verfahrens kann die Selektionskorrektur allerdings auch durch eine simultane Schätzung beider Gleichungen als partielle Maximum Likelihood Schätzung erfolgen, im Folgenden auch SSM(ML). Diese gilt jedoch als noch weniger robust gegenüber Verletzungen der Verfahrensannahmen als die Two-Step-Variante.

3 Empirischer Vergleich

Im Folgenden wird zunächst die Datenbasis und das Analysedesign zum Vergleich der Missing Data Techniken vorgestellt. Dabei wird auf eine konkrete Forschungsfrage Bezug genommen, die jedoch selbst inhaltlich nicht von Interesse ist, sondern nur dem Verfahrensvergleich dient. Danach werden kurz die konkreten Varianten der verwendeten Missing Data Techniken vorgestellt und schließlich die Ergebnisse des Vergleichs präsentiert.

3.1 Datenbasis und Analysedesign

Die Datenbasis für den empirischen Vergleich bildet eine Befragung zur Kombilohnförderung „Mainzer Modell“. Im Rahmen der Evaluation dieser zunächst regional begrenzten, später bundesweit eingesetzten Kombilohnförderung (Kaltenborn et al. 2005, Krug 2009), wurden von TNS Infratest Sozialforschung im Auftrag des Instituts für Arbeitsmarkt- und Berufsforschung (IAB) der Bundesagentur für Arbeit geförderte und ungefördert erwerbstätige Vergleichspersonen befragt. Die Stichprobe umfasste Förderzugänge bzw. Abgänge aus Arbeitslosigkeit im Zeitraum Januar 2001 bis März 2003 (für Details siehe Hartman 2004).

Grundsätzlich stand eine Förderung nach dem Mainzer Modell allen Personen offen, die im Inland eine Beschäftigung aufnehmen dürfen. Es knüpfte nicht an personenbezogenen Merkmalen wie vorheriger Arbeitslosigkeit oder fehlender Qualifikation an. Als Vergleichsgruppe wurden Abgängerinnen bzw. Abgänger aus Arbeitslosigkeit ausgewählt, die keine abgeschlossene Berufsausbildung haben oder langzeitarbeitslos (mindestens seit einem Jahr ununterbrochen arbeitslos gemeldet) waren. Daher werden für die nachfolgende Analyse nur geförderte Personen mit ebensolchen Eigenschaften herangezogen. Die Stichprobe wurde zeitlich und regional proportional zu den Zugängen in die geförderte oder ungeförderte Beschäftigung gezogen. Alle Befragten wurden um die Erlaubnis gebeten, ihre Befragungsdaten mit Prozessdaten verknüpfen zu dürfen. Insgesamt stimmten 74,4 % diesem Anliegen zu. Damit ist die Zustimmungsquote relativ niedrig, verglichen mit der in anderen Erhebungen im Bereich der Arbeitsmarktforschung.

Für den nachfolgenden empirischen Vergleich wird zunächst auf Basis der vollständigen, unbearbeiteten Daten eine Analyse des Einflusses der Kombilohnförderung Mainzer Modell auf die Arbeitszeit der aufgenommenen Beschäftigung durchgeführt[3]:

$$AZEIT_i = \alpha_0 + \beta\ MZM_i + \sum_{l=1}^{4} \gamma_l ALO_{li} + \sum_{j=1}^{h} \delta_j KONT_{ji} + \varepsilon_i \qquad (5)$$

Dabei steht AZEIT für die Arbeitszeit, MZM ist ein binärer Indikator für die Kombilohnförderung im Mainzer Modell und ALO[4] steht für Dummyvariablen (1 bis unter 6 Monate, 6 bis unter 12 Monate, 12 bis unter 24 Monate und ab 24 Monate), welche die klassierte „kumulierte Dauer der Arbeitslosigkeit im Erwerbsleben“ abbilden. Schließlich steht KONT für h diskrete sowie kontinuierliche Kontrollvariablen.

In jedem Datensatz existieren auch ohne die Durchführung von Record Linkage fehlende Werte in einzelnen Variablen. Um die gewählten Missing Data Techniken besser miteinander vergleichen zu können, werden alle Beobachtungen mit fehlenden Werten für die Befragungsvariablen AZEIT, MZM und KONT aus der Analyse entfernt. Dies gewährleistet, dass der einzig relevante Ausfallmechanismus der durch die Zustimmungsverweigerung der Befragten ist. Von den insgesamt 2786 Befragten weisen 2605 keine fehlenden Werte in den Variablen aus der Befragung auf. Von letzteren wiederum stimmen 665 einem Record Linkage nicht zu.

Die Ergebnisse aus dieser Regression mit allen 2605 Beobachtungen stellen die „wahren Werte“ der Koeffizienten dar und dienen daher als Benchmark für den empirischen Vergleich der Missing Data Verfahren, an dem deren Performance gemessen wird. Um die Situation nach einer tatsächlichen Datenverknüpfung zu simulieren (die faktisch jedoch nicht stattfand) werden nun Werte bei den 665 Personen ohne Zustimmung zum Record Linkage gelöscht und im Anschluss die Missing Data Verfahren auf die Daten mit diesen pseudo-fehlenden Werten angewendet. Dabei werden verschiedene Szenarien durchgespielt, wobei schrittweise in immer mehr Variablen fehlende Werte simuliert werden (Tabelle 1).

In Szenario 1 wird angenommen, dass lediglich die abhängige Variable eine Prozessdatenvariable darstellt und damit nur dort fehlende Werte aufgrund der Verknüpfung entstehen. Hierzu werden im Datensatz die 665 Werte der Variable AZEIT für diejenigen Befragten auf „fehlend“ gesetzt, welche ihre Zustimmung zu einer

---

[3] Explizites Ziel im Mainzer Modell war neben der Aktivierung Arbeitsloser zur Aufnahme niedrig entlohnter Beschäftigung vor allem auch die Förderung von Teilzeitbeschäftigung (vgl. Kaltenborn et al. 2005)

[4] Unabhängig von der Zustimmung lagen bei allen Befragten einige wenige Informationen aus den administrativen Daten der Bundesagentur für Arbeit bereits vor (vgl. Hartmann et al. 2002: 173 ff.). Dazu gehört neben der Variablen *Arbeitslosigkeitsdauer* auch die Variable *Stellung im Beruf*. Alle anderen Variablen stammen aus der Befragung, werden aber zum Zweck des Verfahrensvergleiches je nach Szenario wie Prozessdaten mit pseudo-fehlenden Werten versehen.

Datenverknüpfung verweigert hatten. Auf Basis dieses Datensatzes mit pseudo-fehlenden Prozessdaten bei den Personen, die einer Zusammenspielung nicht zugestimmt hatten, werden schließlich Schätzungen der Regressionsgleichung (5) unter Verwendung der in Abschnitt 2 vorgestellten Möglichkeiten des Umgangs mit den fehlenden Werten durchgeführt. Die Ergebnisse der korrigierten Schätzungen können dann mit den tatsächlich auf Basis der vollständigen Daten durchgeführten Schätzungen verglichen werden.

Tabelle 1
Szenarien des Datenausfalls durch zustimmungspflichtiges Record Linkage

Szenario 1	Fehlende Werte in der abhängigen Variablen
Szenario 2	Fehlende Werte in einer unabhängigen Variablen
Szenario 3	Fehlende Werte in mehreren unabhängigen Variablen
Szenario 4	Fehlende Werte in allen unabhängigen Variablen

In Szenario 2 werden ebenfalls in nur einer Variable fehlende Werte simuliert, diese ist nun jedoch mit ALO eine unabhängige Variable der Regressionsanalyse (Gleichung 5). In Szenario 3 werden Ausfälle in der Variablen ALO und einer Reihe weiterer Kontrollvariablen simuliert und in Szenario 4 werden schließlich ALO und alle oben als KONT bezeichneten Kontrollvariablen auf „fehlend“ gesetzt, wenn die Zustimmung zum Record Linkage nicht vorliegt. Natürlich wären noch weitere Szenarien und Abstufungen möglich, die vorliegende Analyse beschränkt sich aufgrund von Platz- und Zeitgründen auf diese vier.

3.2 Implementation der Missing Data Techniken

Die technisch am wenigsten aufwändige Variante ist die Complete Case Analyse. Hier bedarf es keiner gesonderten Schätzverfahren und/oder Software, es werden lediglich alle Fälle mit fehlenden Werten aus der Analyse entfernt, und auf die verbliebenen Fälle werden die üblichen statistischen Schätzverfahren angewendet.

Zur multiplen Imputation der pseudo-fehlenden Prozessdaten wurde in der folgenden Analyse die Sequential Regression Multivariate Imputation (SRMI) Methode verwendet, die in der Software IVEware implementiert ist (Raghunathan et al. 2001; Raghunathan et al. 2002). IVEWare bietet im vorliegenden Anwendungsfall den Vorteil, dass hier komplexe Datenstrukturen berücksichtigt werden können, wie sie in Befragungen auftreten. Neben Imputationsroutinen für kontinuierliche Merkmale bietet IVEWare auch solche für Zähldaten, dichotome und kategoriale Variablen, und es können u. a. Filterbedingungen berücksichtigt werden, z. B. dass nur für Personen mit Kindern die Anzahl dieser Kinder imputiert wird.

Bei der Imputation ist zwischen dem Analysemodell (hier die Variablen in der Regression der Arbeitszeit) und dem Imputationsmodell (für die Imputation verwendeten Variablen) zu unterscheiden. Das Imputationsmodell sollte die Variablen enthalten,

die a) mit der ausfallbelasteten Variable und b) mit dem Ausfall zusammenhängen (Schafer 1997: 143). Daher werden neben den Variablen des Analysemodells im Imputationsmodell noch eine Reihe zusätzlicher Variablen aufgenommen (vgl. Tabelle 7 im Anhang). Es wurden fünf imputierte Datensätze erzeugt, wobei pro Imputation 500 Iterationen durchlaufen wurden.

Zur Schätzung der Parameter im Sample Selection Modell wurde das Stata ado *heckman* verwendet, wobei sowohl die Two-Step als auch die Maximum Likelihood-variante implementiert ist und daher auch beide in Tabelle 2 ausgewiesen wurden. Typischerweise fällt es schwer, geeignete Instrumente zur Durchführung des SSM zu finden. Hier wurden mehrere Variablen in die Selektionsgleichung aufgenommen, von denen allerdings lediglich der Dummy für den Sozialhilfebezug signifikant ist (vgl. Tabelle 7 im Anhang).

3.3 Ergebnisse des empirischen Vergleiches

Tabelle 2 zeigt die Ergebnisse der verschiedenen Missing Data Techniken für die verschiedenen Szenarien im Vergleich zu den Ergebnissen mit den echten, vollständigen Daten. Der Fokus des Vergleichs liegt auf dem Koeffizienten des Dummies für die Kombilohnförderung auf die Arbeitszeit. Darüber hinaus liefert grundsätzlich auch der Vergleich der geschätzten Koeffizienten der pseudo-ausfallbehafteten Variablen mit den „wahren" Koeffizienten Anhaltspunkte zur Bewertung der Missing Data Verfahren. Aufgrund ihres Umfanges wurden die entsprechenden vollständigen Tabellen allerdings in den Anhang verschoben.

Tabelle 2
Koeffizienten und Standardfehler der Dummyvariable „Kombilohnförderung"

OLS [a]- Regressionen der wöchentlichen Arbeitszeit in Stunden	**Benchmark**	**CCA**	**MI**	**SSM (2)**	**SSM (ML)**
	Koeff. (Std.Fehler)	**Koeff. (Std.Fehler)**	**Koeff. (Std.Fehler)**	**Koeff. (Std.Fehler)**	**Koeff. (Std.Fehler)**
Szenario 1	-4,769*** (0,550)	-4,462*** (0,632)	-4,135*** (0,660)	-4,400*** (0,645)	-4,250*** (0,682)
Szenario 2	-4,769*** (0,550)	-4,462*** (0,632)	-4,752*** (0,551)	-4,375*** (0,657)	-4,162*** (0,681)
Szenario 3	-4,769*** (0,550)	-4,462*** (0,632)	-4,349*** (0,674)	-4,346*** (0,675)	-4,107*** (0,665)
Szenario 4	-4,769*** (0,550)	-4,462*** (0,632)	-4,508*** (0,613)	-4,297*** (0,714)	-3,998*** (0,665)

Legende: *p<0,05; **p<0,01;***p<0,001
[a] Ausnahme SSM(ML)

In der mit Benchmark überschriebenen Spalte befindet sich die Schätzung auf Basis der vollständigen Daten. Die Ergebnisse zeigen, dass die Kombilohnförderung einen signifikanten Einfluss auf die Arbeitszeit hatte. Im Mittel arbeiten Geförderte 4,8 Stunden weniger als ungefördert Beschäftigte, dies gilt (natürlich) für alle Szenarien gleichermaßen.

Die Complete Case Analyse zeigt im Vergleich zur Benchmarkanalyse eine leichte Unterschätzung des negativen Einflusses der Förderung. Bedingt durch die geringere Fallzahl zeigt sich auch ein größerer Standardfehler. Dennoch bleibt der Effekt auf dem 0,1 %-Niveau signifikant, so dass sich die aus der Analyse ergebende Schlussfolgerung nicht verändert. Aufgrund des Datanausfallmusters beim Record Linkage (die ausfallbehafteten Beobachtungen sind stets dieselben, unabhängig davon, welche Variablen betrachtet werden) verändert sich die Schätzung über die verschiedenen Szenarien nicht.

Im Gegensatz zur CCA zeigen sich bei den statistischen Korrekturverfahren im engeren Sinne je nach Szenario Unterschiede. Zunächst zur multiplen Imputation: In Szenario 2 mit einer pseudo-ausfallbelasteten Kontrollvariable zeigt sich, dass hier die Schätzung für den Effekt der Förderung deutlich näher an der Benchmarkanalyse liegt als die CCA (bzw. auch das SSM). Dies ändert sich allerdings, wenn statt der unabhängigen Kontrollvariable die abhängige Variable fehlende Werte enthält (Szenario 1). Zum einen steigt der Standardfehler, zum anderen performt die MI schlechter als die CCA bzw. das SSM in der Two-Step Variante und ähnlich gut/schlecht wie die Maximum Likelihood Variante des SSM. In Szenario 3 und 4, wo nicht mehr die abhängige, aber ein großer Teil bzw. alle unabhängigen Variablen fehlende Werte aufweisen, nähert sich der Koeffizient der Dummyvariable „Kombilohnförderung“ wieder dem Benchmark an. Es bleibt jedoch beim hohen Standardfehler.

In der Two-Step Variante des Sample Selection Models zeigt sich eine über alle Szenarien ähnliche Perfomance, etwas unterhalb der der CCA, jedoch mit größeren Standardfehlern. Im Übergang von Szenario 1 zu Szenario 2 zeigen sich im Gegensatz zur MI nur geringe Veränderung der Koeffizienten, die allein dadurch entstehen, dass die Variable ALO in Szenario 2 ausfallbehaftet ist und daher nicht zur Berechnung der Mills Ratio (Selektionsgleichung) verwendet werden kann. Im Vergleich zur MI liefert das Two-Step SSM also bei fehlenden Werten in einer abhängigen Variable ungenauere, in einer unabhängigen Variable genauere Ergebnisse und bei mehreren unabhängigen Variablen in etwa ähnliche Ergebnisse. Dem gegenüber liefert die ML Variante des Sample Selection Models die Koeffizientenschätzungen mit der größten Abweichung zum Benchmark, abgesehen von der MI in Szenario 1, wo die Performance ähnlich ist.

4 Schlussfolgerungen

In der vorliegenden Fallstudie zum Record Linkage wurden die Resultate mehrerer Verfahren beim Umgang mit fehlenden Werten empirisch miteinander verglichen. Der Vorteil des gewählten Analysedesigns besteht darin, dass der Erfolg verschiedener Missing Data Verfahren unter in der Forschungspraxis realistischen Bedingungen analysiert wird, d. h. beim Umgang mit dem zum einen nicht künstlich simulierten sondern empirisch bedingten und zum anderen nicht a priori bekannten sondern grundsätzlich unbekannten (wenn auch wissenschaftlicher Erkenntnis zugäng-

lichen vgl. Hartmann/Krug 2009) Ausfallprozess. Dieser Ausfallprozess ist die versagte Zustimmung zum Rekord Linkage von Prozess- zu Befragungsdaten, wie sie inzwischen in einer Mehrzahl von Befragungen eine Rolle spielt (vgl. z. B. Hethey/Spengler 2009).

Alles in Allem zeichnen die Ergebnisse dieser Fallstudie zum zustimmungsabhängigen Rekord Linkage ein positives Bild der Korrektur von Ausfällen durch Missing Data Techniken. In der Anwendung aller drei Verfahren wird der „wahre“ Koeffizient zwar unterschätzt, das Ausmaß dieser Unterschätzung fällt aber eher gering aus. Zudem ändert sich nichts an den inhaltlichen Schlussfolgerungen, der Koeffizient bleibt stets signifikant neagtiv. Beim Fehlen einer einzigen Variable, welche zudem in der empirischen Analyse lediglich als Kontrollvariable benötigt wird (Szenario 2), zeigt sich die Multiple Imputation als das genaueste Verfahren. Die gute Performance im Vergleich zu anderen Verfahren basiert vermutlich auf der Tatsache, dass die MI als einziges betrachtetes Verfahren auch die tatsächlichen Kovariateninformationen aus den Beobachtungen nutzt, die dem Record Linkage nicht zustimmen. Handelt es sich bei der ausfallbehafteten Variable um die Abhängige einer empirischen Analyse, fällt die Performance der Multiple Imputation im Vergleich zu alternativen Verfahren allerdings ab (Szenario 1). Hier zeigen sich die CCA und das SSM(2) am zuverlässigsten. Fehlen Werte in mehreren oder allen unabhängigen Variablen, liefern mit Ausnahme des SSM(ML) alle betrachteten Verfahren im Grunde ähnliche Ergebnisse (Szenarien 3 und 4).

Literatur

Allmendinger, J.; Kohlmann, A. (2005): Datenverfügbarkeit und Datenzugang am Forschungsdatenzentrum der Bundesagentur für Arbeit im Institut für Arbeitsmarkt- und Berufsforschung. Allgemeines Statistisches Archiv 88: 159-182.

Collins, L.M.; Schafer, J.L.; Kam, C.M. (2001): A comparison of inclusive and restrictive missing-data strategies in modern missing-data procedures. Psychological Methods 6: 330–351.

Engelhardt, H. (1999): Lineare Regression mit Selektion: Möglichkeiten und Grenzen der Heckman-Korrektur. Kölner Zeitschrift für Soziologie und Sozialpsychologie, 51: 706-723.

Graham, J.W. (2008): Missing Data Analysis: Making It Work in the Real World. Annual Review of Psychology, 60: 549-576.

Hartmann, J. (2004): Repräsentative Erhebung zur Evaluation des Mainzer Modells. S. 49-139 in: Gewiese, T.; Hartmann, J.; Krug, G.; Rudolph, H. (Hg.): Das Mainzer Modell aus Sicht der Arbeitnehmer und Betriebe. Befunde aus der Begleitforschung. (Nürnberg: IAB) http://doku.iab.de/externe/2004/k040823w09.pdf (04.05.2009).

Hartmann, J.; Hollederer, A.; Kaltenborn, B.; Rudolph, H.; Vanselow, A.; Weinkopf, C.; Wiedemann, E. (2002): Vom arbeitsmarktpolitischen Sonderprogramm CAST zur bundesweiten Erprobung des Mainzer Modells. 2. Zwischenbericht des Forschungsverbunds "Evaluierung CAST". BMWA-Dokumentation Nr. 516.

Hartmann, J.; Brink, K.; Jäckle, R.; Tschersich, N. (2008): IAB-Haushaltspanel im Niedrigeinkommensbereich - Methoden- und Feldbericht. FDZ Methodenreport 7/2008 (Nürnberg: IAB) http://doku.iab.de/fdz/reporte/2008/MR_07-08.pdf (04.05.2009)

Hartmann, J.; Krug, G. (2009): Verknüpfung von personenbezogenen Prozess- und Befragungsdaten. Selektivität durch fehlende Zustimmung der Befragten? Zeitschrift für ArbeitsmarktForschung 42(2).

Heckman, J.J. (1979): Sample Selection Bias as a Specification Error. Econometrica 47(1): 153-162.

Hethey, T.; Spengler, A. (2009): Matching process generated business data and survey data. The case of KombiFiD in Germany. FDZ Methodenreport 01/2009. http://doku.iab.de/fdz/reporte/2009/MR_01-09.pdf (04.05.2009).

Kaltenborn, B.; Krug, G.; Rudolph, H.; Weinkopf, C.; Wiedemann, E. (2005): Evaluierung der arbeitsmarktpolitischen Sonderprogramme CAST und Mainzer Modell. Bundesministerium für Wirtschaft und Arbeit. Forschungsbericht Nr. 552, Berlin.

Krug, G. (2009): In-work benefits for low-wage jobs. Can additional income reduce employment stability? European Sociological Review (im Erscheinen).

Puhani, P.A. (2000): The Heckman Correction for Sample Selection and Its Critique. Journal of Economic Surveys 14(1): 53-68.

Raghunathan, T.E.; Lopkowski, J.M.; van Hoeweyk, J.; Solenberger, P. (2001): A Multivariate Technique for Multiply Imputing Missing Values Using a Series of Regression Models. Survey Methodology 27: 85-96.

Raghunathan, T.E.; Solenberger, P.; van Hoeweyk, P. (2002): IVEware: Imputation and Variance Estimation Software. User Guide. http://www.isr.umich.edu/src/smp/ive (04.05.2009).

Rässler, S. (2002): Statistical Matching. A Frequentist Theory, Practical Applications and Alternative Bayesian Approaches. New York, Berlin: Springer.

Ridder, G. (1992): An empirical evaluation of some models for non-random attrition in panel data. Structural Change and Economic Dynamics 3(2): 337-355.

Rubin, D.B. (1976): Inference with missing data (with discussion). Biometrica 63: 581-592.

Rubin, D.B. (1987): Multiple Imputation for Nonresponse in Surveys. New York: Wiley.

Schafer, J.L. (1997): Analysis of Incomplete Multivariate Data. London, u. a.: Chapman & Hall/CRC.

Weins, C. (2006): Multiple Imputation. S. 205-216 in: Behnke, J.; Gschwend, T.; Schindler, D.; Schnapp, K.-U. (Hg): Methoden der Politikwissenschaft. Neuere quantitative Methoden Analyseverfahren. Baden-Baden: Nomos.

Wirth, H.; Müller, W. (2004): Mikrodaten der amtlichen Statistik als eine Datengrundlage der empirischen Sozialforschung. S. 93-127 in: Diekmann, A. (Hg.): Methoden der Sozialforschung (Sonderheft 44 der Kölner Zeitschrift für Soziologie und Sozialpsychologie). Wiesbaden: VS Verlag für Sozialwissenschaften.

Anhang

Tabelle 3
Performance der CCA (alle Szenarien) - OLS-Regressionen der wöchentlichen Arbeitszeit in Stunden

	Benchmark	Szenario 1-4
	Koeff. (Std.Fehler)	**Koeff. (Std.Fehler)**
Kombilohnförderung: ja	-4,769***	-4,462***
	(0,550)	(0,632)
Kumulierte Dauer der Arbeitslosigkeit (R: unter einem Monat)[a]		
Ein bis unter sechs Monate	-0,730	-1,156
	(0,591)	(0,667)
Sechs bis unter zwölf Monate	-1,178	-1,475*
	(0,656)	(0,737)
Zwölf bis unter vierundzwanzig Monate	-2,165**	-2,030**
	(0,682)	(0,777)
Über vierundzwanzig Monate	-1,685*	-2,222*
	(0,818)	(0,923)
Alter	0,282	0,110
	(0,173)	(0,197)
Alter (quadr.)	-0,004	-0,002
	(0,002)	(0,003)
Stellung im Beruf (R: Arbeiter)		
Angestellter	-3,212***	-3,874***
	(0,456)	(0,520)
Sonst.	-2,476*	-3,456**
	(1,152)	(1,339)
Branche (R: Leih-/Zeitarbeitsfirma)		
Reinigungsgewerbe	-9,196***	-7,989***
	(0,940)	(1,086)
Hotel- und Gaststättengewerbe	-2,694**	-2,741*
	(0,961)	(1,095)
Callcenter	-4,573***	-3,898**
	(1,202)	(1,367)
Energie- und Wasser; Bergbau, Verarb. Gewerbe	0,102	-0,104
	(0,774)	(0,867)
Handel	-4,319***	-4,455***
	(0,739)	(0,827)
Verkehr/Nachrichtenüberm., Banken, Versicherungen Öffentlicher Dienst/Sozialversicherung	-4,230***	-3,883***
	(0,870)	(0,981)
Bereich andere Dienstleistungen, k. A.	-2,996***	-2,677***
	(0,674)	(0,757)
Haushaltskontext (R: alleinstehend)		
alleinerziehend	-2,211**	-2,388*
	(0,823)	(0,950)
Nicht ewt.partner, ohne Kind	1,424	1,205
	(0,879)	(1,031)

	Benchmark	Szenario 1-4
	Koeff. (Std.Fehler)	**Koeff. (Std.Fehler)**
Ewt. partner, ohne Kind	2,351** (0,892)	2,135* (1,020)
Nicht ewt. partner, Kind	3,204*** (0,881)	3,461*** (1,010)
Ewt. partner, Kind	-1,260 (0,905)	-1,639 (1,038)
Zahl der Kinder (unter 18 Jahren im Haushalt)	0,101 (0,293)	0,174 (0,332)
Nebentätigkeit vorhanden	-3,452*** (1,009)	-3,324** (1,156)
Einkommen sehr wichtig	0,333 (0,432)	0,375 (0,491)
Arbeit sehr wichtig	0,954* (0,437)	0,958 (0,499)
Freizeit sehr wichtig	-0,242 (0,483)	0,264 (0,545)
Familie sehr wichtig	-1,080* (0,532)	-1,486* (0,608)
Region: Ost	4,718*** (0,444)	4,598*** (0,505)
Art der Kündigung letzte Beschäftigung (R: sonstiges)		
Arbeitnehmerkündigung	1,558 (0,825)	1,545 (0,909)
Arbeitgeberkündigung	1,423** (0,453)	1,645** (0,519)
Konstante	33,678*** (3,027)	36,712*** (3,479)
Adjusted R^2	0,2125	0,2305
N	2605	1940

Legende: *p<0,05; **p<0,01;***p<0,001
Grau unterlegt: Koeffizienten ausfallbelasteter Variablen

Tabelle 4
Performance der MI nach Szenarien - OLS-Regressionen der wöchentlichen Arbeitszeit in Stunden

	Benchmark	Szenario 1	Szenario 2	Szenario 3	Szenario 4
	Koeff. (Std.Fehler)	**Koeff. (Std.Fehler)**	**Koeff. (Std.Fehler)**	**Koeff. (Std.Fehler)**	**Koeff. (Std.Fehler)**
Kombilohnförderung: ja	-4,769***	-4,135***	-4,752***	-4,349***	-4,508***
	(0,550)	(0,660)	(0,551)	(0,674)	(0,613)
Kumulierte Dauer der Arbeitslosigkeit (R: unter einem Monat)[a]					
Ein bis unter sechs Monate	-0,730	-0,929	-0,916	-0,880	-0,793
	(0,591)	(0,598)	(0,671)	(0,592)	(0,598)
Sechs bis unter zwölf Monate	-1,178	-1,199	-1,504*	-1,252	-1,189
	(0,656)	(0,692)	(0,725)	(0,658)	(0,678)
Zwölf bis unter vierundzwanzig Monate	-2,165**	-1,708*	-2,351**	-2,244**	-2,138**
	(0,682)	(0,700)	(0,801)	(0,688)	(0,707)
Über vierundzwanzig Monate	-1,685*	-1,737*	-2,309*	-1,926*	-1,808*
	(0,818)	(0,826)	(0,913)	(0,820)	(0,843)
Alter	0,282	0,206	0,272	0,054	0,114
	(0,173)	(0,197)	(0,173)	(0,181)	(0,210)
Alter (quadr.)	-0,004	-0,003	-0,004	-0,001	-0,002
	(0,002)	(0,003)	(0,002)	(0,002)	(0,003)
Stellung im Beruf (R: Arbeiter)					
Angestellter	-3,212***	-3,867***	-3,213***	-4,050***	-4,117***
	(0,456)	(0,523)	(0,456)	(0,509)	(0,498)
Sonst.	-2,476*	-2,964	-2,581*	-1,922	-1,185
	(1,152)	(1,526)	(1,152)	(1,888)	(1,482)
Branche (R: Leih-/Zeitarbeitsfirma)					
Reinigungsgewerbe	-9,196***	-8,087***	-9,121***	-7,891***	-7,682***
	(0,940)	(1,194)	(0,942)	(1,050)	(1,162)
Hotel- und Gaststättengewerbe	-2,694**	-2,853**	-2,650**	-2,859*	-2,890**
	(0,961)	(1,087)	(0,962)	(1,206)	(1,111)
Callcenter	-4,573***	-3,343**	-4,543***	-4,136**	-3,817**
	(1,202)	(1,265)	(1,203)	(1,329)	(1,240)
Energie- und Wasser; Bergbau, Verarb. Gewerbe	0,102	-0,183	0,134	0,345	0,321
	(0,774)	(0,809)	(0,774)	(1,142)	(0,832)
Handel	-4,319***	-4,298***	-4,251***	-4,313***	-4,445***
	(0,739)	(0,789)	(0,740)	(0,796)	(0,905)
Verkehr/Nachrichtenüberm., Banken, Versicherungen Öffentlicher Dienst/ Sozialversicherung	-4,230***	-3,734***	-4,189***	-3,597**	-3,560***
	(0,870)	(0,914)	(0,871)	(1,153)	(1,027)

	Benchmark	Szenario 1	Szenario 2	Szenario 3	Szenario 4
	Koeff. (Std.Fehler)	**Koeff. (Std.Fehler)**	**Koeff. (Std.Fehler)**	**Koeff. (Std.Fehler)**	**Koeff. (Std.Fehler)**
Bereich andere Dienst-leistungen, k. A.	-2,996***	-2,543***	-2,962***	-2,609***	-2,694**
	(0,674)	(0,705)	(0,674)	(0,743)	(0,842)
Haushaltskontext (R: alleinstehend)					
alleinerziehend	-2,211**	-2,366**	-2,195**	-2,117*	-1,822
	(0,823)	(0,870)	(0,826)	(0,906)	(0,944)
Nicht ewt.partner, ohne Kind	1,424	1,353	1,417	0,906	1,162
	(0,879)	(0,931)	(0,880)	(1,040)	(1,118)
Ewt. partner, ohne Kind	2,351**	2,322*	2,332**	2,250	2,622*
	(0,892)	(0,969)	(0,892)	(1,169)	(1,329)
Nicht ewt. partner, Kind	3,204***	3,414***	3,224***	3,603***	3,471**
	(0,881)	(0,979)	(0,883)	(1,065)	(1,228)
Ewt. partner, Kind	-1,260	-1,405	-1,254	-1,430	-1,597
	(0,905)	(0,936)	(0,907)	(0,951)	(1,176)
Zahl der Kinder (unter 18 Jahren im Haushalt)	0,101	0,007	0,100	0,030	0,114
	(0,293)	(0,310)	(0,293)	(0,398)	(0,421)
Nebentätigkeit vorhanden	-3,452***	-2,462*	-3,461***	-3,066**	-3,481**
	(1,009)	(1,009)	(1,012)	(1,176)	(1,081)
Einkommen sehr wichtig	0,333	0,361	0,365	0,305	0,430
	(0,432)	(0,428)	(0,432)	(0,440)	(0,470)
Arbeit sehr wichtig	0,954*	1,016*	0,941*	1,089*	0,970*
	(0,437)	(0,442)	(0,438)	(0,447)	(0,458)
Freizeit sehr wichtig	-0,242	0,010	-0,256	-0,180	-0,029
	(0,483)	(0,574)	(0,483)	(0,485)	(0,503)
Familie sehr wichtig	-1,080*	-1,314*	-1,073*	-1,216*	-1,624**
	(0,532)	(0,611)	(0,532)	(0,545)	(0,555)
Region: Ost	4,718***	4,439***	4,772***	4,729***	4,734***
	(0,444)	(0,471)	(0,446)	(0,447)	(0,522)
Art der Kündigung letzte Beschäftigung (R: sonstiges)					
Arbeitnehmerkündigung	1,558	1,766*	1,485	1,453	1,015
	(0,825)	(0,847)	(0,826)	(0,826)	(0,801)
Arbeitgeberkündigung	1,423**	1,865***	1,353**	1,400**	1,752***
	(0,453)	(0,454)	(0,455)	(0,460)	(0,509)
Konstante	33,678***	34,674***	33,988***	37,564***	36,552***
	(3,027)	(3,553)	(3,034)	(3,134)	(3,353)
Adjusted R^2	0,2125	n.v.	n.v.	n.v.	n.v.
N	2605	2605	2605	2605	2605

Legende: *p<0,05; **p<0,01;***p<0,001 ; n.v.: nicht verfügbar;
Anmerkung: Standardfehler basieren auf Rubin 1987
Grau unterlegt: Koeffizienten ausfallbelasteter Variablen

Tabelle 5
Performance des SSM(2) nach Szenarien - OLS-Regressionen der wöchentlichen Arbeitszeit in Stunden

	Benchmark	Szenario 1	Szenario 2	Szenario 3	Szenario 4
	Koeff. (Std.Fehler)	**Koeff. (Std.Fehler)**	**Koeff. (Std.Fehler)**	**Koeff. (Std.Fehler)**	**Koeff. (Std.Fehler)**
Kombilohnförderung: ja	-4,769***	-4,400***	-4,375***	-4,346***	-4,297***
	(0,550)	(0,645)	(0,657)	(0,675)	(0,714)
Kumulierte Dauer der Arbeitslosigkeit (R: unter einem Monat)[a]					
Ein bis unter sechs Monate	-0,730	-1,213	-1,153	-1,156	-1,152
	(0,591)	(0,677)	(0,661)	(0,661)	(0,661)
Sechs bis unter zwölf Monate	-1,178	-1,488*	-1,474*	-1,476*	-1,473*
	(0,656)	(0,739)	(0,732)	(0,731)	(0,732)
Zwölf bis unter vierundzwanzig Monate	-2,165**	-2,194**	-2,024**	-2,025**	-2,019**
	(0,682)	(0,845)	(0,771)	(0,771)	(0,771)
Über vierundzwanzig Monate	-1,685*	-2,271*	-2,210*	-2,213*	-2,204*
	(0,818)	(0,929)	(0,916)	(0,916)	(0,916)
Alter	0,282	0,098	0,094	0,106	0,105
	(0,173)	(0,199)	(0,200)	(0,195)	(0,195)
Alter (quadr.)	-0,004	-0,001	-0,001	-0,002	-0,002
	(0,002)	(0,003)	(0,003)	(0,003)	(0,003)
Stellung im Beruf (R: Arbeiter)					
Angestellter	-3,212***	-3,927***	-3,934***	-3,875***	-3,874***
	(0,456)	(0,532)	(0,535)	(0,516)	(0,516)
Sonst.	-2,476*	-3,411*	-3,417*	-3,438**	-3,439**
	(1,152)	(1,343)	(1,343)	(1,328)	(1,328)
Branche (R: Leih-/Zeitarbeitsfirma)					
Reinigungsgewerbe	-9,196***	-8,280***	-8,301***	-7,994***	-7,993***
	(0,940)	(1,233)	(1,257)	(1,078)	(1,078)
Hotel- und Gaststättengewerbe	-2,694**	-2,899*	-2,903*	-2,728*	-2,731*
	(0,961)	(1,141)	(1,145)	(1,087)	(1,087)
Callcenter	-4,573***	-4,119**	-4,139**	-3,922**	-3,923**
	(1,202)	(1,438)	(1,453)	(1,357)	(1,357)
Energie- und Wasser; Bergbau, Verarb. Gewerbe	0,102	-0,157	-0,165	-0,110	-0,107
	(0,774)	(0,875)	(0,878)	(0,861)	(0,861)
Handel	-4,319***	-4,544***	-4,549***	-4,474***	-4,472***
	(0,739)	(0,847)	(0,850)	(0,821)	(0,821)

	Benchmark	Szenario 1	Szenario 2	Szenario 3	Szenario 4
	Koeff. (Std.Fehler)	**Koeff. (Std.Fehler)**	**Koeff. (Std.Fehler)**	**Koeff. (Std.Fehler)**	**Koeff. (Std.Fehler)**
Verkehr/Nachrichten-überm., Banken, Versicherungen Öffentlicher Dienst/ Sozialversicherung	-4,230***	-4,012***	-4,015***	-3,894***	-3,893***
	(0,870)	(1,016)	(1,019)	(0,974)	(0,973)
Bereich andere Dienst-leistungen, k. A.	-2,996***	-2,860***	-2,871***	-2,686***	-2,684***
	(0,674)	(0,843)	(0,855)	(0,752)	(0,751)
Haushaltskontext (R: alleinstehend)					
alleinerziehend	-2,211**	-2,039	-2,021	-2,350*	-2,354*
	(0,823)	(1,180)	(1,206)	(0,945)	(0,945)
Nicht ewt.partner, ohne Kind	1,424	1,223	1,236	1,232	1,230
	(0,879)	(1,032)	(1,033)	(1,024)	(1,023)
Ewt. partner, ohne Kind	2,351**	2,381*	2,389*	2,148*	2,151*
	(0,892)	(1,133)	(1,144)	(1,012)	(1,012)
Nicht ewt. partner, Kind	3,204***	3,842**	3,857**	3,497***	3,496***
	(0,881)	(1,266)	(1,290)	(1,004)	(1,004)
Ewt. partner, Kind	-1,260	-1,326	-1,320	-1,644	-1,641
	(0,905)	(1,213)	(1,224)	(1,029)	(1,029)
Zahl der Kinder (unter 18 Jahren im Haushalt)	0,101	0,174	0,171	0,180	0,179
	(0,293)	(0,333)	(0,333)	(0,330)	(0,330)
Nebentätigkeit vorhanden	-3,452***	-3,353**	-3,351**	-3,340**	-3,335**
	(1,009)	(1,158)	(1,159)	(1,147)	(1,147)
Einkommen sehr wichtig	0,333	0,307	0,315	0,318	0,373
	(0,432)	(0,511)	(0,507)	(0,505)	(0,487)
Arbeit sehr wichtig	0,954*	1,109	1,112	1,080	0,960
	(0,437)	(0,584)	(0,590)	(0,560)	(0,495)
Freizeit sehr wichtig	-0,242	0,335	0,328	0,321	0,264
	(0,483)	(0,564)	(0,561)	(0,558)	(0,540)
Familie sehr wichtig	-1,080*	-1,482*	-1,477*	-1,423*	-1,487*
	(0,532)	(0,608)	(0,609)	(0,622)	(0,603)
Region: Ost	4,718***	4,767***	4,738***	4,730***	4,583***
	(0,444)	(0,609)	(0,580)	(0,574)	(0,502)
Art der Kündigung letzte Beschäftigung (R: sonstiges)					
Arbeitnehmerkündigung	1,558	1,753	1,512	1,747	1,509
	(0,825)	(1,003)	(0,904)	(1,002)	(0,905)
Arbeitgeberkündigung	1,423**	1,626**	1,629**	1,638**	1,629**
	(0,453)	(0,521)	(0,515)	(0,519)	(0,515)
Konstante	33,678***	35,533***	35,542***	35,495***	35,535***
	(3,027)	(4,208)	(4,213)	(4,275)	(4,188)
rho		0,251	0,256	0,238	0,263
		(n.v.)	(n.v.)	(n.v.)	(n.v.)

	Benchmark	Szenario 1	Szenario 2	Szenario 3	Szenario 4
	Koeff. (Std.Fehler)	**Koeff. (Std.Fehler)**	**Koeff. (Std.Fehler)**	**Koeff. (Std.Fehler)**	**Koeff. (Std.Fehler)**
lambda		2,508 (5,026)	2,566 (5,196)	2,379 (-4,929)	2,633 (-5,301)
Adjusted R^2	0,2125	n.v.	n.v.	n.v.	n.v.
N	2605	2605	2605	2605	2605

Legende: *p<0,05; **p<0,01;***p<0,001 ; n.v.: nicht verfügbar; robuste Standardfehler
Grau unterlegt: Koeffizienten ausfallbelasteter Variablen

Tabelle 6
Performance des SSM(ML) nach Szenarien - ML-Schätzung der wöchentlichen Arbeitszeit in Stunden

	Benchmark	Szenario 1	Szenario 2	Szenario 3	Szenario 4
	Koeff. (Std.Fehler)	Koeff. (Std.Fehler)	Koeff. (Std.Fehler)	Koeff. (Std.Fehler)	Koeff. (Std.Fehler)
Kombilohnförderung: ja	-4,769***	-4,250***	-4,162***	-4,107***	-3,998***
	(0,550)	(0,682)	(0,681)	(0,665)	(0,665)
Kumulierte Dauer der Arbeitslosigkeit (R: unter einem Monat)[a]					
Ein bis unter sechs Monate	-0,730	-1,336	-1,102	-1,130	-1,111
	(0,591)	(0,721)	(0,655)	(0,656)	(0,656)
Sechs bis unter zwölf Monate	-1,178	-1,552	-1,516*	-1,522*	-1,524*
	(0,656)	(0,799)	(0,726)	(0,726)	(0,727)
Zwölf bis unter vierundzwanzig Monate	-2,165**	-2,562**	-1,971**	-1,987**	-1,960*
	(0,682)	(0,840)	(0,765)	(0,765)	(0,765)
Über vierundzwanzig Monate	-1,685*	-2,359*	-2,112*	-2,121*	-2,105*
	(0,818)	(0,998)	(0,912)	(0,911)	(0,911)
Alter	0,282	0,066	0,053	0,077	0,074
	(0,173)	(0,213)	(0,212)	(0,194)	(0,194)
Alter (quadr.)	-0,004	-0,001	-0,001	-0,001	-0,001
	(0,002)	(0,003)	(0,003)	(0,003)	(0,003)
Stellung im Beruf (R: Arbeiter)					
Angestellter	-3,212***	-4,032***	-4,049***	-3,856***	-3,851***
	(0,456)	(0,562)	(0,561)	(0,513)	(0,513)
Sonst.	-2,476*	-3,567*	-3,571*	-3,977**	-4,009**
	(1,152)	(1,441)	(1,440)	(1,330)	(1,333)
Branche (R: Leih-/Zeitarbeitsfirma)					
Reinigungsgewerbe	-9,196***	-8,939***	-8,985***	-8,060***	-8,051***
	(0,940)	(1,173)	(1,173)	(1,073)	(1,074)
Hotel- und Gaststättengewerbe	-2,694**	-3,387**	-3,383**	-2,969**	-2,991**
	(0,961)	(1,186)	(1,185)	(1,083)	(1,083)
Callcenter	-4,573***	-4,572**	-4,623**	-3,887**	-3,909**
	(1,202)	(1,477)	(1,477)	(1,351)	(1,347)
Energie- und Wasser; Bergbau, Verarb. Gewerbe	0,102	-0,215	-0,245	0,047	0,043
	(0,774)	(0,940)	(0,939)	(0,858)	(0,857)
Handel	-4,319***	-4,768***	-4,781***	-4,609***	-4,631***
	(0,739)	(0,897)	(0,897)	(0,818)	(0,818)

	Benchmark	Szenario 1	Szenario 2	Szenario 3	Szenario 4
	Koeff. (Std.Fehler)	**Koeff. (Std.Fehler)**	**Koeff. (Std.Fehler)**	**Koeff. (Std.Fehler)**	**Koeff. (Std.Fehler)**
Verkehr/Nachrichten-überm., Banken, Versicherungen Öffentlicher Dienst/ Sozialversicherung	-4,230***	-4,386***	-4,381***	-4,109***	-4,129***
	(0,870)	(1,063)	(1,063)	(0,967)	(0,967)
Bereich andere Dienst-leistungen, k. A.	-2,996***	-3,343***	-3,367***	-2,903***	-2,898***
	(0,674)	(0,824)	(0,824)	(0,752)	(0,751)
Haushaltskontext (R: alleinstehend)					
alleinerziehend	-2,211**	-1,166	-1,137	-2,026*	-2,019*
	(0,823)	(1,032)	(1,033)	(0,941)	(0,942)
Nicht ewt.partner, ohne Kkind	1,424	1,192	1,231	1,145	1,116
	(0,879)	(1,105)	(1,104)	(1,017)	(1,016)
Ewt. partner, ohne Kind	2,351**	2,855**	2,859**	2,022*	1,996*
	(0,892)	(1,102)	(1,101)	(1,005)	(1,004)
Nicht ewt. partner, Kind	3,204***	4,663***	4,685***	3,482***	3,478***
	(0,881)	(1,095)	(1,095)	(0,995)	(0,995)
Ewt. partner, Kind	-1,260	-0,646	-0,644	-1,649	-1,625
	(0,905)	(1,124)	(1,123)	(1,027)	(1,027)
Zahl der Kinder (unter 18 Jahren im Haushalt)	0,101	0,161	0,152	0,169	0,154
	(0,293)	(0,360)	(0,360)	(0,330)	(0,330)
Nebentätigkeit vorhanden	-3,452***	-3,424**	-3,408**	-3,417**	-3,392**
	(1,009)	(1,246)	(1,245)	(1,138)	(1,137)
Einkommen sehr wichtig	0,333	0,151	0,182	0,184	0,348
	(0,432)	(0,531)	(0,531)	(0,529)	(0,485)
Arbeit sehr wichtig	0,954*	1,430**	1,429**	1,340*	0,957
	(0,437)	(0,541)	(0,541)	(0,537)	(0,492)
Freizeit sehr wichtig	-0,242	0,518	0,483	0,467	0,330
	(0,483)	(0,590)	(0,590)	(0,587)	(0,537)
Familie sehr wichtig	-1,080*	-1,479*	-1,467*	-1,293*	-1,523*
	(0,532)	(0,655)	(0,655)	(0,650)	(0,599)
Region: Ost	4,718***	5,172***	5,062***	5,058***	4,608***
	(0,444)	(0,549)	(0,544)	(0,540)	(0,499)
Art der Kündigung letzte Beschäftigung (R: sonstiges)					
Arbeitnehmerkündigung	1,558	2,234*	1,482	2,232*	1,483
	(0,825)	(0,993)	(0,891)	(0,989)	(0,892)
Arbeitgeberkündigung	1,423**	1,609**	1,676**	1,647**	1,685***
	(0,453)	(0,559)	(0,512)	(0,556)	(0,512)
Konstante	33,678***	33,143***	33,263***	33,290***	33,805***
	(3,027)	(3,763)	(3,753)	(3,454)	(3,444)
athrho		0,880***	0,876***	0,849***	0,865***
		(0,108)	(0,108)	(0,107)	(0,104)

	Benchmark	Szenario 1	Szenario 2	Szenario 3	Szenario 4
	Koeff. (Std.Fehler)	**Koeff. (Std.Fehler)**	**Koeff. (Std.Fehler)**	**Koeff. (Std.Fehler)**	**Koeff. (Std.Fehler)**
lnsigma		2,417*** (0,029)	2,417*** (0,030)	2,412*** (0,030)	2,417*** (0,029)
Adjusted R^2	0,2125	n.v.	n.v.	n.v.	n.v.
N	2605	2605	2605	2605	2605

Legende: *p<0,05; **p<0,01;***p<0,001 ; n.v.: nicht verfügbar; robuste Standardfehler
Grau unterlegt: Koeffizienten ausfallbelasteter Variablen

Tabelle 7
Zusätzliche Variablen

im Imputationsmodell	in der Selektionsgleichung (Instrumente)
Von ABM in Beschäftigung	Suchanstrengungen
Von and. Maßnahme in Beschäftigung	Bereits einmal Stelle wegen zu niedrigem Einkommen abgelehnt
	Wie wurde die Stelle gefunden (Arbeitsvermittlung; Bekannte/ Freunde; Eigene Initiative; sonstiges)
Vorher Weiterbildung ja	
Lohnersatzleistungen vom Arbeitsamt (Arbeitslosengeld; Arbeitslosenhilfe; Keine)	Sozialhilfebezug vor Beschäftigungsaufnahme
Beschäftigung befristet	
Bruttolohn in Euro (falls Angabe)	
Bruttolohn: Angabe nicht verweigert	
Geschlecht weiblich	
Zufrieden mit Lohn: ja	
Zeit in (geförderter) Beschäftigung bis Interviewzeitpunkt	
Beschäftigung zum Interviewzeitpunkt beendet?	
Nationalität deutsch	
Berufserfahrung in Jahren	
Nettohaushaltseinkommen	

In dieser Reihe sind zuletzt erschienen

Nr.	Autor(en)	Titel	Datum
38/2008	Kruppe, Th. Rudloff, K.	Wirksamkeit beruflicher Weiterbildungsmaß-nahmen: Eine mirkoökonometrische Evaluation der Ergänzung durch das ESF-BA-Programm in der Zeit von 2000 bis 2002 auf Basis von Prozessdaten der Bundesagentur für Arbeit	9/08
39/2008	Brixy, U.	Welche Betriebe werden verlagert: Beweggrün-de und Bedeutung von Betriebsverlagerungen	10/08
40/2008	Oberschachtsiek, D.	Founders' Experience and Self-Employment Duration : The Importance of Being a 'Jack-of-all-Trades'. An Analysis Based on Competing Risks	10/08
41/2008	Kropp, P. Schwengler, B.	Abgrenzung von Wirtschaftsräumen auf der Grundlage von Pendlerverflechtungen : Ein Methodenvergleich	10/08
42/2008	Krug, G. Popp, S.	Soziale Herkunft und Bildungsziele von Jugend-lichen im Armutsbereich	12/08
43/2008	Hofmann, B.	Work Incentives? Ex-Post Effects of Unemploy-ment Insurance Sanctions : Evidence from West Germany	12/08
44/2008	Büttner, Th. Rässler, S.	Multiple Imputation of Right-Censored Wages in the German IAB Employment Sample Consider-ing Heteroscedasticity	12/08
1/2009	Bruckmeier, K. Schwengler, B.	The Impact of federal social policies on spatial income inequalities in Germany : Empirical evi-dence from social security data	1/09
2/2009	Büttner, Th. Jacobebbing-haus, P. Ludsteck, J.	Occupational Upgrading and the Business Cycle in West Germany	2/09
3/2009	Donado, A. Wälde, K.	Trade Unions Go Global!	1/09
4/2009	Schanne, N. Weyh, A.	What makes start-ups out of unemployment dif-ferent?	1/09
5/2009	Trappmann, M. Christoph, B. Achatz, J. Wenzig, C. Müller, G. Gebhardt, D.	Design and stratification of PASS : A New Panel Study for Research on Long Term Unemploy-ment	3/09
6/2009	Ruppe, K.	Eingliederungszuschüsse und Betriebszugehö-rigkeitsdauer in Westdeutschland	4/09

Stand: 12.05.2009

Eine vollständige Liste aller erschienen IAB-Discussion Paper finden Sie unter
http://www.iab.de/de/publikationen/discussionpaper.aspx

Impressum

IAB-Discussion Paper 7/2009

Herausgeber
Institut für Arbeitsmarkt- und Berufs-
forschung der Bundesagentur für Arbeit
Regensburger Str. 104
90478 Nürnberg

Redaktion
Regina Stoll, Jutta Palm-Nowak

Technische Herstellung
Jutta Sebald

Website
http://www.iab.de

Bezugsmöglichkeit
http://doku.iab.de/discussionpapers/2009/dp0709.pdf

Rückfragen zum Inhalt an:

Gerhard Krug
Telefon 0911.179 3387
E-mail gerhard.krug@iab.de