Forschungsdatenzentrum
BIBB.FDZ
im Bundesinstitut für Berufsbildung

BIBB-FDZ
Daten- und Methodenberichte

Nr. 2 / 2011

Imputation fehlender Werte für die Einkommensvariable in der BIBB/BAuA-Erwerbstätigenbefragung 2006

Holger Alda, Daniela Rohrbach-Schmidt

Bundesinstitut für Berufsbildung
- Forschungsdatenzentrum -
Robert-Schuman-Platz 3
53175 Bonn

Telefon: 0228 / 107 - 20 41
Fax: 0228 / 107 - 20 20
E-Mail: fdz@bibb.de

www.bibb-fdz.de

Imputation fehlender Werte für die Einkommensvariable der BIBB/BAuA-Erwerbstätigenbefragung 2006

Autoren: Dr. Holger Alda, Dr. Daniela Rohrbach-Schmidt[1]

Abstract

Fehlende Angaben in Umfragedaten reduzieren nicht nur die Anzahl der verwertbaren Beobachtungen in einem Datensatz, sondern können unter bestimmten Umständen mit einem Rückgang der Validität der Daten, einem Verlust an Präzision und gegebenenfalls mit einer Verzerrung der Analyseergebnisse verbunden sein. Die Literatur kennt verschiedene Möglichkeiten, dem Problem fehlender Werte zu begegnen. Eine angemessene Herangehensweise an das Problem fehlender Werte umfasst dabei erstens die Identifizierung des zugrundeliegenden Mechanismus und die Beachtung der Konsequenzen für die statistische Inferenz, sowie zweitens die Anwendung geeigneter Verfahren zum Umgang mit den fehlenden Werten vor der eigentlichen Auswertung. Im vorliegenden Papier wird die Problematik fehlender Werte bei einzelnen Items für die Forschungspraxis und gängige Strategien zum Umgang mit diesen am Beispiel der Einkommensvariable in der BIBB/BAuA-Erwerbstätigenbefragung 2006 diskutiert.

[1] Forschungsdatenzentrum im Bundesinstitut für Berufsbildung (BIBB-FDZ), Robert-Schuman-Platz 3, D-53175 Bonn, Deutschland. E-Mail: alda@bibb.de ; rohrbach@bibb.de.

Inhaltsverzeichnis

Hinweis zur Zählweise bei Versionsnummern

Änderungen gegenüber der Vorversion ohne größere inhaltliche Relevanz werden durch fortlaufende Nummern *nach* dem Punkt dokumentiert (zweite Ebene). Inhaltlich relevante Änderungen führen demgegenüber zu einer fortlaufenden Nummerierung auf der ersten Ebene.

1 Einleitung

Unvollständige Daten können - etwa durch den Rückgang der Zahl analysierbarer Fälle - bereits in deskriptiven Analysen ein mehr oder minder beachtenswertes Problem darstellen, insbesondere bei der Auswertung von Teilgesamtheiten. Wenn die fehlenden Werte nicht zufällig zustande kommen, ist damit unter Umständen ein Rückgang der Validität, ein Verlust an Präzision und somit eine Verzerrung der Analyseergebnisse verbunden. Für valide Ergebnisinterpretationen kann es demnach hilfreich sein, eine sorgfältige (”sensitivity“) Analyse der Mechanismen durchzuführen, die den fehlenden Werten zugrunde liegen, und sich der Implikationen bewusst zu sein, die sich für die statistische Inferenz ergeben. Gegebenenfalls sind auf der Grundlage einer solchen Analyse Verfahren anzuwenden, mit denen sich fehlende Daten durch plausible Werte ersetzen (imputieren) lassen.

Ziel des vorliegenden Papiers ist es, die Problematik fehlender Werte für die Forschungspraxis und gängige Strategien zum Umgang mit diesen zu skizzieren. Die Thematik wird auf einen konkreten und häufigen Praxisfall – fehlende Werte bei der Einkommensangabe in Umfragen – angewendet. Als Datengrundlage für unsere Analysen dient die BIBB/BAuA-Erwerbstätigenbefragung 2006.

Das zweite Kapitel skizziert mögliche Problematiken bei der Analyse unvollständiger Daten. Hierzu werden zentrale Mechanismen illustriert, die fehlenden Werten zugrunde liegen und auf Konsequenzen für die statistische Inferenz eingegangen (Abschnitt 2.1). Anschließend werden Verfahren zum Umgang mit fehlenden Werten dargestellt (Abschnitt 2.2). Das dritte Kapitel wendet sich dann unserer Anwendung zu, wo die fehlenden Einkommensangaben in der BIBB/BAuA-Erhebung ersetzt werden. Das machen wir auf zweifache Art und Weise. Zum einen ersetzen wir nur die fehlenden Werte, zum anderen imputieren wir zusätzlich die Ausreißerwerte der Lohnverteilung am oberen und unteren Rand. Die so erweiterten beiden Lohnvariablen sind in den Scientific-Use-File der BIBB/BAuA-Erhebung integriert. Das vorliegende Papier dokumentiert demnach die Entstehung und die Konstruktionsprinzipien der beiden zusätzlichen Variablen. Der Beitrag schließt - neben einer Zusammenfassung - mit einem Ausblick auf Erweiterungen der diskutierten Methoden und weitere Anwendungen auf (BIBB-)Datensätze (Kapitel 4).

2 Die Analyse unvollständiger Daten

2.1 Mechanismen fehlender Werte und ihre Implikationen

In der Praxis der Umfrageforschung ist zwischen dem sogenannten unit-non-response, also fehlenden Angaben für ausgewählte Erhebungseinheiten und item-non-response – fehlenden Werten auf einzelnen Variablen von Erhebungseinheiten, die grundsätzlich zur Teilnahme an der Befragung bereit sind - zu unterscheiden[2]. Zu den wesentlichen Ursachen

[2] Im Fall von Dropouts bei Panelerhebungen sind beide Fälle unter Umständen nicht mehr so eindeutig unterscheidbar. Verfahren fehlender Werte bei Panelbefragungen oder anderen hierarchisch

für das Fehlen ganzer Fälle gehören etwa die Nichterreichbarkeit von Zielpersonen und die explizite Teilnahmeverweigerung an der Umfrage.

Sofern sich ex-post[3] bei dem Vergleich der Stichprobendaten mit zuverlässigen Informationen über die Merkmalsverteilungen in der Grundgesamtheit eine Abweichung zeigt, wird dem Problem des Unit-Nonresponse in der Regel durch eine Zellgewichtung begegnet (zu Gewichtungsproblemen in der Datenanalyse siehe etwa Gabler 2004). Das bedeutet, dass durch die Multiplikation mit einem Faktor größer oder kleiner eins die Stichprobendaten dem Profil der Grundgesamtheit angepasst werden[4]. Im Unterschied dazu treten (folgenreiche) fehlende Werte bei einzelnen Variablen insbesondere dann auf, wenn Befragte die Frage nicht beantworten können oder wollen. So kann es sein, dass die Frage nicht verstanden wird oder aus Sicht des Befragten schlicht nicht sinnvoll zu beantworten ist. Bei Berücksichtigung der wichtigsten Regeln guter Frageformulierung und -reihenfolge, einem erfolgreichen Pretest und anderen Instrumentarien der Umfragepraxis ist der wohl häufigste Grund für das Auftreten fehlender Fälle bei einzelnen Variablen die bewusste Verweigerung der Herausgabe von Informationen. Dies tritt insbesondere bei sozial brisanten oder die Privatsphäre betreffenden, sensiblen Fragen auf. Eine der Variablen, die in der Regel immer einen beachtenswerten Anteil fehlender Werte aufweist, ist die Einkommensvariable (vgl. etwa Rässler 2000). Es ist zudem bekannt, dass Personen mit sehr hohen Einkommen die Beantwortung von Einkommensfragen oft verweigern.

Das Vorliegen bzw. Nichtvorliegen von Werten bei einer sensiblen Variable ist also zumeist nicht völlig zufällig. Dies hat Konsequenzen für die Zulässigkeit des Schließens von der Stichprobe auf die Grundgesamtheit. Insofern, als die fehlenden Daten eigentlich Daten sind, die wir gerne beobachtet hätten, sollten wir den Mechanismus berücksichtigen, nach dem die eigentlich zu beobachtenden Einheiten zu unbeobachteten geworden sind. Denn der Auswahlprozess umfasst im Fall fehlender Werte nicht nur den Auswahlprozess der beobachteten Einheiten, sondern auch den Auswahlprozess der unbeobachteten.

Fehlende Daten können die Eigenschaften der Schätzer und die Teststatistik beeinflussen. Entscheidend dafür, ob und wie stark die Zulässigkeit für statistische Inferenz eingeschränkt wird, ist die Art und Weise, wie die Wahrscheinlichkeit für das Auftreten eines fehlenden Wertes von anderen beobachteten oder nicht beobachteten Variablen und seinem eigenen Wert abhängt. Formal ausgedrückt ist der Mechanismus fehlender Werte also die Wahrscheinlichkeit *Pr*, dass ein Set an einzelnen Werten *r* – bedingt durch die beobachteten und unbeobachteten Fälle Y_o und Y_m – fehlt:

(2.1) $Pr(r \mid y_o, y_m)$.

---

strukturierten Daten diskutieren beispielsweise Verbeke und Molenberghs (2000) und Molenberghs und Verbeke (2005).

[3] Ex-ante wird in der Regel mit Hilfe verschiedener Strategien (die Auswahl des geeigneten Erhebungsinstruments, Incentives, postalischen An- und Erinnerungsschreiben etc.) versucht, das Auftreten von unit-non-response zu minimieren.

[4] Man spricht deshalb bei der Zellgewichtung auch von Redressment oder Anpassungsgewichtung.

Mit Blick auf die zugrundeliegenden Mechanismen der fehlenden Daten[5] werden nach Rubin (1976) und Little und Rubin (1987)[6] die folgenden Unterscheidungen vorgenommen.

1. *Missing Completely at Random (MCAR):* Die Wahrscheinlichkeit für das Fehlen einer Messung ist unabhängig von unbeobachteten und beobachteten Werten, d.h. es gilt:

(2.2) $Pr(r \mid y_o, y_m) = Pr(r)$

2. *Missing at Random (MAR*): Die Wahrscheinlichkeit für das Fehlen einer Messung ist konditional unabhängig von den unbeobachteten Werten, gegeben die Werte der beobachteten Daten, d.h. es gilt:

(2.3) $Pr(r \mid y_o, y_m) = Pr(r \mid y_o)$

3. *Missing Not at Random (MNAR):* Die Wahrscheinlichkeit für das Fehlen einer Messung ist abhängig von den unbeobachteten Werten, d.h. 2.1 lässt sich nicht vereinfachen!

Unter *MCAR*-Bedingungen kann unabhängig davon, ob Verfahren der Häufigkeits-, Wahrscheinlichkeits- oder Bayesstatistik angewendet werden, der Prozess ignoriert werden, der die fehlenden Werte generiert (Molenberghs und Verbeke, 2005: 487). Eine fehlende Einkommensangabe ist unter *MCAR*-Bedingung beispielweise also weder von der Höhe des (unbeobachteten) Einkommens, noch von der (beobachteten Werten für) Bildung, dem Geschlecht und dem Alter abhängig, sondern kommt etwa durch die unbewusst erzeugte Unleserlichkeit der Eintragung des Befragten oder durch Fehler bei der Dateneingabe zustande. Eine Möglichkeit, die *MCAR*-Annahme zu testen, ist es, die Beobachtungen für die relevante Variable *j* in zwei Gruppen zu teilen und zwar (1) solche Fälle, für die gültige Werte auf einer oder mehreren anderen Variablen vorliegen und (2) solche Fälle, für die keine gültigen Werte auf der einen oder mehreren anderen Variablen vorliegen. Für den Fall, dass *MCAR* gegeben ist, sollten beide Gruppen eine Zufallsstichprobe derselben Grundgesamtheit mit gleichen Verteilungsparametern darstellen (vgl. Little und Rubin 1987, zitiert nach Verbeke und Mohlenberghs 2000: 223).

Unter *MAR*-Bedingungen ist der Umgang mit fehlenden Werten im Vergleich zu MNAR-Bedingungen (s. u.) zwar erheblich einfacher, aber die Begründung für das Vorliegen von *MAR* im Gegensatz zu *MNAR* kann in der Praxis problematisch werden. Angewendet auf das Beispiel Einkommen würde unter *MAR*-Bedingungen eine fehlende Einkommensangabe also nicht von der Höhe des (unbeobachteten) Einkommens, aber von der (beobachteten) Bildung abhängen. Im Fall von *MAR* ist es demnach wichtig, diejenigen Kovariaten zu identifizieren, die den *MAR*-Mechanismus der unbeobachteten Daten determinieren.

[5] Die Taxonomie basiert auf der Faktorisierung der Gesamtverteilung der Daten in einem Faktor für den Messprozess, und einem Faktor für den Missingnessprozess, bedingt durch die Outcomes. Das ist die Grundlage für Selektionsmodelle.
[6] Vgl. die Darstellung in Verbeke und Mohlenberghs (2000).

Unter *MNAR*-Bedingungen (auch ”informative dropout“) kann Inferenz nur unter Zuhilfenahme von weiteren Annahmen erfolgen, über die die beobachteten Daten allein keine Information enthalten (vgl. Molenberghs und Verbeke 2005: 486). Da fehlende Einkommensangaben unter *MNAR*-Bedingungen also etwa sowohl von der Höhe des (unbeobachteten) Einkommens, als auch von der (beobachteten) Bildung abhängig sind, sind die Annahmen im Idealfall auf externe Informationen zu stützen. Denn selbst wenn alle beobachtbaren Informationen berücksichtigt werden, ist die Ursache für den fehlenden Wert immer noch abhängig von den unbeobachteten Werten selbst. Um zu validen Schlüssen zu gelangen, ist demnach ein gemeinsames annahmegestütztes Modell für den Daten- und den Missingnessprozess erforderlich. Von zentraler Wichtigkeit in der Analyse von Datensätzen mit *MNAR* gilt es sorgfältig zu untersuchen, inwieweit sich die Ergebnisse je nach den eingebrachten Annahmen unterscheiden. Einfache ”ad-hoc“-Methoden, wie sie im ersten Teil des nächsten Abschnitts kurz diskutiert werden, sollten bei *MNAR* in der Praxis keine Verwendung finden (vgl. etwa Molenberghs und Verbeke 2005: 490).

2.2 Verfahren im Umgang mit fehlenden Werten

2.2.1 Einfache Verfahren: CC, LOCF, MI, einfache Regressionsimputation

Vor der Beschäftigung mit der Frage, welches Verfahren im Umgang mit fehlenden Werten geeignet ist, gilt es, den zugrundeliegenden Prozess für die fehlenden Werte zu bestimmen. Nur für den (relativ seltenen) *MCAR*-Fall führen die folgenden einfachen Verfahren zu validen Ergebnissen: Complete Case Analysis (CC), Last Observation Carried Forward (LOCF, bei Paneldaten) und verschiedene Arten der Mean Imputation (MI).

Die einfachste Methode im Umgang mit fehlenden Werten ist die Analyse nur derjenigen Fälle im Datensatz, für die vollständige Informationen (CC) vorliegen. Alle Analysen und Inferenzstatistiken beruhen damit auf derselben Fallauswahl. Problematisch an dieser Vorgehensweise kann sein, dass vielleicht ein nicht unbeträchtlicher Anteil an Informationen - bedingt durch sinkende Fallzahlen - verloren geht. Die Alternative zur Löschung unvollständiger Fälle ist die einfache Ersetzung der fehlenden Werte durch plausible geschätzte Werte. Zur Ermittlung der geschätzen Werte kann auf Information desselben Falls (Panel), ähnliche Fälle (Mittelwertimputation) oder beide Arten von Fällen (z.B. conditional mean imputation, hot deck imputation) zurückgegriffen werden.

Bei Panelerhebungen kann es in seltenen Fällen plausibel sein, den fehlenden Wert durch den letzten beobachteten Wert zu substituieren. Dabei nimmt man (in der Regel unrealistischerweise) an, dass seit der letzten Messung (etwa der Weiterbildungsaktivität eines Unternehmens) keine Veränderung seit dem Ausfall der Erhebungseinheit eingetreten ist. Bei der einfachen Mittelwertimputation werden die fehlenden Werte durch den Mittelwert

der anderen Einheiten bei der gleichen Variable ersetzt. Bei der konditionalen MI werden die fehlenden Werte durch die Mittelwerte in Subgruppen (gebildet nach Ausprägung einer beobachteten Variablen) ersetzt. Ein anderes Verfahren ist die Ersetzung der fehlenden durch vorhergesagte Werte eines multivariaten Regressionsmodells auf Basis der beobachteten Werte ersetzt. Problematisch hieran ist, dass durch die Ersetzung von Werten, die auf der Regressionslinie liegen die Stärke der Korrelation künstlich erhöht wird. Bei der Hot-deck-Imputation findet eine Ersetzung durch beobachtete Werte eines möglichst ähnlichen Falls im Datensatz (Cold-deck: Ersetzen durch möglichst ähnlichen Fall in anderem Datensatz) statt. Für alle diese einfachen Imputationsverfahren[7] gilt, dass diese unter *MCAR* zwar konsistente Punktschätzer ergeben können, die Präzision aber bei einem anderen Ausfallmechanismus als *MCAR* überschätzt wird, wenn tatsächlich gemessene und imputierte Werte gleichgewichtig in die Analyse eingehen.

2.2.2 Mehrstufige, multiple Imputationsverfahren

Die oben genannten Probleme einfacher Ersetzungsverfahren lassen sich durch die Verwendung mehrstufiger und multipler Imputationsverfahren minimieren. Anstelle der Ersetzung durch plausible (vorhergesagte, sonst wie ermittelte) Werte wird bei mehrstufigen Verfahren dem ermittelten Wert ein in einem Zufallsmechanismus generierter Wert α hinzu addiert. Um an dieser Stelle nicht alle möglichen Verfahren diskutieren zu müssen, beschränken wir uns auf den später auch anzuwendenden Fall der multiplen Imputation einer intervallskalierten sowie zensierten Variable[8]. Gartner (2005) macht einen Vorschlag für den Umgang mit der Zensierung der Lohnangaben in den prozessproduzierten IAB-Personendaten, woran wir unsere eigenen Ansatz und die entsprechenden Berechnungen anlehnen. Das Vorgehen erläutern wir in diesem Abschnitt kurz theoretisch, bevor wir das Verfahren im dritten Abschnitt auf die Daten der BIBB/BAuA-Erwerbstätigenbefragung für das Jahres 2006 anwenden.

Die Einkommensfunktion ist – wie es in größeren Stichproben von Personen eigentlich immer der Fall sein sollte - für alle Einheiten der Stichprobe (loglinear) normal verteilt[9].

---

[7] Neben den erwähnten sind die Ersetzung durch Experteneinschätzung und logische Imputation zu nennen.

[8] Die Anwendung eines Zensierungsverfahrens wird möglich, weil Antwortverweigerer der Einkommensangabe um die Information gebeten wurden, ob sie mehr oder weniger als 1500 Euro im Monat verdienen. Darüber hinaus gehen wir davon aus, dass wir durch diese Angabe eine externe Information für die Berücksichtigung eines eventuellen MNAR-Prozesses haben, denn wir haben durch die Angabe zumindest eine grobe Information über die eigentlich unbeobachtete Lohnhöhe.

[9] Einkommen bedeutet in diesem Papier immer Einkommen aus Erwerbsarbeit. Streng genommen sind „nur“ die Löhne und Gehälter von abhängig Beschäftigten in größeren Personenstichproben (annähernd) loglinear normal verteilt (oft mit einer gewissen Linksschiefe). Auf Besonderheiten bei Selbstständigen, Freiberuflern und sozialversicherungsrechtlich besonderen Beschäftigtengruppen wie etwa 400-Euro-Kräften wird im Verlauf der Analysen eingegangen.

Deshalb können wir für die fehlenden (verweigerten) Einkommensangaben W_{miss} annehmen, dass die gute Annäherung an den tatsächlichen, aber nicht beobachtbaren Einkommenswert die Addition eines Störterms ε_i mit der Standardabweichung σ zu dem zu einer Regressionsgleichung hervorgehenden Vorhersagewert $x_i'\hat{\beta}$ ist und sich σ aus der tatsächlichen (beobachtbaren) Verteilung der Einkommensvariable ergibt. Zur Berücksichtigung eines (eventuellen) *MNAR*-Ausfallprozesses nutzen wir den Vorteil aus, dass in unserem Fall der fehlende Einkommenswert in der BIBB-Erwerbstätigenbefragung 2006 zwar abhängig vom unbeobachteten ist, wir aber die zusätzliche („externe“) Information haben, ob dieser unbeobachtete Wert über oder unter 1500 Euro liegt.

Dies wiederum bedeutet, dass wir das Problem fehlender Werte über einen Imputationsprozess aus einer abgeschnittenen (truncated) Verteilung lösen, indem wir aus einer Standardnormalverteilung den Zufallswert ε_i so ziehen, dass $x_i\beta + \varepsilon_i$ entweder größer oder kleiner als 1500,- Euro (gleich α) ist. Gartner (2005: 4) zeigt, dass wir die konditionalen Werte für ε_i über die Inverse der abgeschnittenen Normalverteilungsfunktion $G(\varepsilon)\varepsilon \rightarrow Y$ mit $Y \in [0,1]$ aufgelöst nach ε_i wie folgt ermitteln können:

(2.4) $\Phi^{-1} (Y (1-\Phi(\alpha)) + \Phi(\alpha)) = \varepsilon$.

Demzufolge berechnen wir die fehlenden Werte W_i^{imp} als

(2.5) $\ln W_i^{imp} = \varepsilon_i\hat{\sigma} + x_i'\hat{\beta}$.

Little und Rubin (1987) schlagen (darüber hinaus) vor, dass der Imputationsalgorhythmus mehrfach und in einer Bayesianischen Vorgehensweise wiederholt wird. Rässler/ Gartner (2005) realisieren das in Form eines Markov-Chain-Monte-Carlo-Algorhythmus (MCMCA) für die soeben beschriebene Vorgehensweise bei einer zensierten Variable.

Der Hintergrund der Anwendung multipler Imputationen ist - wie im vorigen Abschnitt angesprochen - das bei einem ”frequentistischen“ Ansatz (der einfachen Imputation) das dem Datenausfall zugrundeliegende Identifikationsproblem ignoriert werde. Bei einer einzigen Ergänzung gehen tatsächlich gemessene und imputierte Werte gleichgewichtig ein. Werden dagegen plausible Werte *m*-fach imputiert, spiegeln die m Datensätze die (unbekannte) Variablilität der Schätzwerte wider. Um an dieser Stelle keinen zu hohen Aufwand zu betreiben und für konventionelle Anwendungen interessant zu bleiben, haben wir auf die Anwendung eines MCMCA verzichtet und gehen davon aus, dass wir einerseits über die Zensierungsinformation und andererseits über 1000 Wiederholungen und die Verwendung des Mittelwertes der imputierten Werte für jede Erhebungseinheit das Identifikationsproblem lösen, das dem Datenausfall zugrunde liegt. Dieses konkrete Vorgehen ist auch der Aufnahme der imputierten Lohnvariable(n) in den SUF geschuldet, denn es wird *ein* Datensatz angeboten, und nicht m Datensätze, wo für jedes Untersuchungsergebnis dann zunächst der Mittelwert aus den *m* Einzelergebnissen gebildet werden müsste.

3 Fehlende Einkommensangaben in der BIBB/BAuA-Erwerbstätigenbefragung 2006

3.1 Datenerhebung und Ziele

Die BIBB/BAUA-Erwerbstätigenbefragung 2006 ist eine telefonische, computerunterstützte Repräsentativbefragung von 20.000 Erwerbstätigen in Deutschland, die gemeinsam vom Bundesinstitut für Berufsbildung (BIBB) und von der Bundesanstalt für Arbeitsschutz und Arbeitsmedizin (BAuA) durchgeführt und vom Bundesministerium für Bildung und Forschung (BMBF) gefördert wird. Ziel der Erhebung ist es, differenzierte repräsentative Informationen über Erwerbstätige und Arbeitsplätze in Deutschland für Forschungsfragen der quantitativen Berufs- und Qualifikationsforschung sowie der Arbeitsschutzberichterstattung bereit zu stellen. Im Mittelpunkt der Befragung stehen zum einen Fragen zum Arbeitsplatz (Tätigkeitsschwerpunkte, Anforderungsniveau, Kenntnisanforderungen, Arbeitsanforderungen, Weiterbildungsbedarf, Arbeitsbedingungen, Arbeitsbelastungen etc.), zum anderen wird der Zusammenhang zwischen Bildung und Beschäftigung thematisiert (Schul-, Aus- und Weiterbildung, Berufsverlauf, ausbildungsadäquate Beschäftigung, Berufswechsel, Verwertbarkeit beruflicher Qualifikationen, etc.). Verschiedene Berufssystematiken erlauben dabei eine differenzierte Darstellung nach Erwerbs- und Ausbildungsberufen.

3.2 Die (fehlenden) Werte der Einkommensvariable

In der BIBB/BAUA-Erwerbstätigenbefragung 2006 werden Personen nach ihrem monatlichen Bruttoverdienst[10] gefragt (f518). Für den Fall, dass die befragte Person die Antwort verweigert oder mit ”weiß nicht“ beantwortet hat, wurde diese stattdessen gebeten anzugeben, ob der monatliche Bruttoverdienst weniger als 1500 Euro beträgt (f519)[11]. In Tabelle 1 befinden sich die zusammenfassenden Statistiken für die *gültigen* Einkommensangaben.

[10] Der genaue Wortlaut ist: ”Nun zu ihrem Bruttoverdienst. Kindergeld rechnen Sie bitte nicht mit. Wie hoch ist ihr monatlicher Bruttoverdienst aus Ihrer Tätigkeit als …“. Für Selbstständige und Freiberufler wurde ergänzt: „Gemeint ist nicht der Geschäftsumsatz oder –gewinn“, bei freier Mitarbeit: „gemeint ist nicht der Umsatz“ und bei allen übrigen: „d.h. Lohn und Gehalt nach Abzug der Steuern und Sozialversicherung“.

[11] Der Fragetext lautet: „Würden Sie mir dann vielleicht sagen: Beträgt Ihr monatlicher Bruttoverdienst weniger als 1500 Euro?“

Tabelle 1: Verteilungsmaße der Einkommensvariable in der BIBB/BAuA-Erwerbstätigenbefragung 2006 (Basis: gültige Einkommensangabe; Angabe in Euro)

Verteilungsmaß	Ursprungswerte	logarithmiert
Mittelwert	2.600	7,603
Standardabweichung	2.154	7,675
10-Perzentil	700	6,551
25-Perzentil	1.500	7,313
50-Perzentil	2.300	7,741
75-Perzentil	3.300	8,102
90-Perzentil	4.500	8,412
gültige Werte	16.956	16.956

Quelle: BIBB/BAuA-Erwerbstätigenbefragung 2006

Bei insgesamt 20 Tsd. realisierten Interviews gibt es demzufolge 3.044 fehlende Werte. In Anlehnung an das zweite Kapitel interessiert uns vor allem die Einkommensverteilung, denn diese möchten wir möglichst auch mit den imputierten Werten erhalten. Ohne weitere Informationen wäre es jetzt schwierig, etwas über den Ausfallmechanismus zu sagen, etwa ob eher Personen mit höherem oder niedrigerem Einkommen die Lohnangabe verweigern. An dieser Stelle ist die Information aus der nachfolgenden Frage sehr wertvoll, die nur an die entsprechenden Antwortverweigerer gestellt wird. Mit dem Wissen, ob diese Personen mehr oder weniger als 1500 Euro monatlich verdienen, lassen sich zumindest einige grobe Anhaltspunkte zum Ausfallmechanismus darstellen.

Leider ist es so, dass eine gewisse Anzahl[12] an Antwortverweigerern der Lohnangabe auch keine Antwort auf die Frage gibt, ob sie mehr oder weniger als 1500 Euro verdienen. Auch das ließe sich wieder durch einen eigenen mehr oder minder anspruchsvollen Imputationsansatz lösen. Zur Vermeidung eines übermäßigen Aufwands wählen wir eine Vereinfachung. Wir ersetzen die fehlenden Angaben der ”Doppelverweigerer“ durch den Vorhersagewert (ohne Störterm-Addition) einer logistischen Regression mit einigen wenigen Kovariaten[13]. Zumindest verschieben sich durch diese Imputation nicht die Proportionen (Prozentanteile) der imputierten und nicht imputierten Zensierungsvariable. Konkret bedeutet dies, dass in beiden Fällen je etwa 21 Prozent der Befragungseinheiten angeben, weniger als 1500 Euro und demnach etwa 79 Prozent angeben, dass sie mehr als 1500 Euro monatlich verdienen.

[12] 357 Personen (von 3044 gleich etwa elf Prozent) haben sowohl zu f518 als auch f519 keine Angaben gemacht.

[13] Das sind: berufliche Stellung, wöchentliche Arbeitszeit (in sozialversicherungsrechtlich oder sonst wie relevanten Kategorien), einer einfachen Variable für den Migrationshintergrund, ob zur Arbeit in relevanter Weise gependelt wird (ja/nein) sowie der Familienstand. Demnach modellieren wir einen *MAR*-Prozess. Die Ergebnisse der konditionalen Bedingungen hierfür befinden sich im Anhang in Tabelle A1.

3.3 Identifizierung des zugrundeliegenden Mechanismus und seine Implikationen

Bevor wir uns mit dem Imputationsalgorhythmus beschäftigen, überprüfen wir mit einer logistischen Regressionsanalyse die *MAR*- und *MCAR*-Annahme (siehe Tabelle A2 im Anhang). Da wir keine Rangordnung für die uns interessierenden kategorialen Ausprägungen bilden können[14] und die Information nicht verschenken möchten, dass Personen, die angeben, weniger als 1500 Euro zu verdienen, in jedem Fall nicht mehr als 1500 Euro verdienen (bzw. eine gültige Lohnangabe haben), berechnen wir ein multinominales Logit-Modell, in der die Logit-Koeffizienten für Nicht-Verweigerer auf null gesetzt sind und diese Personengruppe demnach die Basiskategorie bzw. Referenzgruppe (’baseline‘) bilden. Nach den entsprechenden empirischen Ergebnissen laut der Tabelle im Anhang A2 ist die *MCAR*-Annahme abzulehnen, denn es gibt sowohl im unteren und oberen Lohnbereich bei den Antwortausfällen signifikante Effekte. So verweigern am unteren Rand der Lohnverteilung Männer seltener die Angabe ihres Bruttoeinkommens als am oberen Rand (Frauen vice versa). Überhaupt lässt sich feststellen, dass fast ausschließlich umgedrehte Vorzeichen für die beiden Verweigerergruppen berechnet werden, d.h. wenn eine unabhängige Variable am unteren Rand der Lohnverteilung einen ’positiven’ Einfluss auf die Antwortverweigerung hat, verweigern Personen mit dem gleichen Merkmal am oberen Rand der Lohnverteilung seltener die Angabe ihres Einkommens ist, relativ zu Personen mit gültigen Einkommensangaben. Darüber hinaus sind die Ergebnisse der Logit-Analyse zu berücksichtigen, wenn wir die fehlenden Angaben mittels eines Regressionsverfahrens schätzen. Notwendig wird ein zweistufiges Verfahren, wo in der ersten Stufe die Selektionskontrolle erfolgt (verweigert ja/nein) und in der zweiten die eigentlich interessierende Lohnhöhe geschätzt wird. Mit anderen Worten: aus den Ergebnissen der Logit-Analyse lässt sich ableiten, dass einfache Ersetzungsverfahren nur bedingt zufriedenstellende Resultate erbringen, denn nach den empirischen Analyseergebnissen muss die *MCAR*-Annahme bei der vorliegenden Datensatzstruktur verworfen werden.

Wie schon angedeutet, ist eine Besonderheit des Datensatzes, dass mit der Variable f519 eine binäre Zusatzinformation über die Links- bzw. Rechtszensierung der fehlenden Einkommensvariable (f518) vorliegt. Diese Zusatzinformation werden wir nicht nur für den Imputationsalgorhytmus nutzen, sondern begründen damit auch, warum wir keinen MNAR-Prozess modellieren (müssen). Für die Imputation schätzen wir mit einer Tobitregression die vermutlichen Werte der Einkommensverweigerer und addieren einen gleich verteilten Zufallswerte α. Im Zuge der Addition der vorhergesagten Werte mit α wird demnach die Information über die Zensierung der tatsächlichen (aber nicht bekannten) Werte nach oben bzw. unten einbezogen. Das führt im Ergebnis dazu, dass die ergänzten Daten auf der Einkommensvariable f518 nicht widersprüchlich zu den gemachten Angaben in f519 sind.

[14] In der abhängigen kategorialen Variable gibt es drei Ausprägungen auf zwei Dimensionen (Einkommen verweigert ja/ nein, und wenn nein, ob höher oder niedriger als 1500 Euro).

Im ersten Schritt bilden wir eine Variable *missing*, die für gültige Fälle auf der Variable f518 den Wert 0, für fehlende Werte auf f518 und einem Einkommen über 1500 Euro (f519=2) den Wert 1 und für fehlende Werte auf f518 und einem Einkommen unter 1500 Euro (f519=1) den Wert -1 annimmt. Aufgrund der Tatsache, dass im Bereich der sehr hohen und sehr niedrigen Einkommen (0,5 und 99,5 Perzentil der Lohnverteilung im gesamten Datensatz) vielleicht eher unplausible Ausreißerwerte enthalten sind, führen wir das Verfahren zweifach durch, in dem wir in einer eigenständigen Analyse diese Ausreißerwerte ebenfalls den Kategorien 1 bzw. -1 auf der Indikatorvariable *missing* entsprechend zuordnen. Die folgende Häufigkeitstabelle 2 gibt die Werte der Variable *missing* für beide Vorgehensweisen wieder.

Tabelle 2: Häufigkeiten gültiger und verweigerter Einkommensangaben

Kategorie für Variable *missing*	**Häufigkeit** nur fehlende Werte	**Häufigkeit** fehlende Werte und Ausreißer*
gültige Einkommensangabe	16.956	16.819
nicht gültig, weniger als 1500 Euro	651	714
nicht gültig, 1500Euro und mehr	2.393	2.467
insgesamt	20.000	20.000

* unterstes und oberstes 0,5-Perzentil der originalen Verteilung

Anmerkung: Personen, die weder eine gültige Einkommensangabe haben noch f519 beantworten (n=357), werden auf Basis des linearen Vorhersagewertes einer logistischen Regression den beiden „nicht gültig“-Kategorien zugeordnet.

Quelle: BIBB/BAuA-Erwerbstätigenbefragung 2006

Im nächsten Schritt sagen wir mit Hilfe einer Tobitregression die logarithmierten[15] Löhne durch gängige, d.h. in der Literatur bekannte Einkommensdeterminanten vorher. Sie lassen sich grob in qualifikatorische, betriebliche und soziodemografische Personenmerkmale einteilen. Vor dem Hintergrund der Tatsache, dass die Wirkung der Kovariaten auf das Einkommen zwischen Männern und Frauen sowie zwischen West- und Ostdeutschen systematisch variiert, werden die Regressionen getrennt für die vier Gruppen berechnet. Die Ergebnisse der Regressionsanalysen geben wir auf Anfrage gerne weiter.

[15] Ein zusätzlicher Nutzen dieser Transformation der abhängigen Variablen ist, dass die Regressionskoeffizienten annähernd als prozentuale Änderung interpretiert werden können.

3.4 Anwendung eines Imputationsverfahrens

Auf Basis der ermittelten vorhergesagten Werte wird getrennt für die vier Gruppen von Personen und den beiden jeweiligen Ausfallgruppen (missing = 1 und missing = -1) ein α-Wert anhand der folgenden Gleichungen generiert:

$$(3.4) \qquad a(1) = \frac{\ln(1500) - x_i\tilde{\beta}}{\hat{\sigma}}$$

$$(3.5) \qquad a(-1) = \frac{-x_i\hat{\beta}}{\hat{\sigma}}$$

Da *ln(0)* in Gleichung 3.5 nicht definiert wäre, wurde die Untergrenze (null Euro Verdienst) symbolisch auf einen Euro angehoben. Die Verteilungen von σ ergeben sich aus den vier Teilgleichungen der Regressionsanalysen für die einzelnen Personengruppen (Männer/Frauen, Ost/West). Anschließend werden jeweils für die vier Gruppen von Personen und jeweils beide Ausfallgruppen (*missyes* = 1 und *missyes* = -1) die Einkommensangaben durch die Addition der α-Werte mit den vorhergesagten Werten wie folgt imputiert:

(3.6) $\ln(wage_{imp}) = x_i'\tilde{\beta} + \alpha_{i1}$

(3.7) $\ln(wage_{imp}) = x_i'\tilde{\beta} + \alpha_{i-1}$.

Da die Addition der *α*-Werte für eine einzelne Einheit mehr oder minder zufällig erfolgt, wiederholen wir die Ersetzungen insgesamt 1000mal, speichern jedes einzelne Ergebnis ab und bestimmen den ”tatsächlichen“ Wert auf der Personenebene als Mittelwert aus diesen 1000 Wiederholungen[16].

3.5 Originale und gebildete Einkommensvariable im Vergleich

In diesem Abschnitt berichten wir über die Ergebnisse unserer Vorgehensweise. Zuerst vergleichen wir deskriptiv die beiden imputierten Verteilungen der Einkommensvariablen mit der für die nicht imputierten Werte. Anschließend stellen wir Regressionskoeffizienten (gleich Lohnertragsraten) für gängige Einkommensdeterminanten gegenüber, wenn wir die unterschiedlichen Einkommensverteilungen als abhängige Variable mit einem Regressionsmodell unterlegen. Schließlich erfolgt eine kleine deskriptive Übung zu Reichweite und Gültigkeit, wenn Forschende die imputierte Einkommensvariable verwenden. Zuerst stellt Tabelle 3 die Ergebnisse der Imputationsprozeduren zusammen. Es handelt sich um gängige Streuungsmaße für beide Imputationsarten (ersetzen nur von fehlenden sowie von fehlenden und Ausreißer-Werten) im Vergleich zur Ursprungsvariable.

[16] Damit legen wir fest, dass der Mittelwert der Additionen des Störterms *auf der Ebene einzelner Personen* die Verteilung der 1000 Einzelwerte zufrieden stellend beschreibt. Das ist eine gute Festlegung, wenn die 1000 Einzelwerte etwa annähernd normal verteilt (und dabei nicht sonderlich links- oder rechtsschief) sind.

Tabelle 3: Vergleich imputierter und nicht imputierter Werte (alle Angaben in Euro)

Variable	Mittelwert (Standardabw.)	Quotient Perzentil 90/10	Quotient Perzentil 80/20	gültige Fälle
Originalwerte				
Ursprungsvariable	2600 (2155)	6,429	2,917	16.954
Imputation 1*	2648 (2045)	5,625	2,769	20.000
Imputation 2**	2571 (1542)	5,625	2,720	20.000
logarithmiert				
Ursprungsvariable	7,603 (0,816)	1,284	1,151	16.954
Imputation 1	7,642 (0,789)	1,258	1,142	20.000
Imputation 2	7,642 (0,722)	1,258	1,140	20.000

* Imputation fehlender Werte ** Imputation fehlender Werte und von Ausreißern

Quelle: BIBB/BAuA-Erwerbstätigenbefragung 2006

Ein erster interessanter Effekt ist das Lohnniveau in der Datensatzstruktur vor und nach den beiden Imputationsroutinen. Werden nur die fehlenden Werte ersetzt, steigt das Lohniveau im Datensatz geringfügig. Ersetzt man zusätzlich die Werte für die Ausreißer, nimmt das Lohnniveau hingegen um einen ähnlichen Betrag ab[17]. Dies bedeutet, dass bei der Ersetzung der fehlenden Werte und der Ausreißerwerte insbesondere relativ hohe Einkommen etwas geschmälert werden, jedenfalls mehr, als kleinere Einkommen durch die gleiche Imputationsroutine angehoben werden.

Etwas deutlicher sind die Auswirkungen der jeweiligen Imputationen auf die Streuungsmaße. Da es sich im Kern bei der Anwendung unseres Imputationsansatzes jeweils um Glättungen der Verteilung der Einkommensvariable handelt, sinken die Standardabweichungen, wie bei fast jeder Anwendung von Imputationsverfahren. Der Effekt ist umso größer, je mehr originale Werte durch „geglättete“ Werte ersetzt werden (hier der Ausreißer-Werte). Während sich der Quotient des 90/10-Perzentils zwischen beiden Imputationsarten nicht unterscheidet, ist bei dem für das 80/20-Perzentil ein kleinerer Unterschied festzustellen. Demzufolge wird bei dem Imputationsverfahren, dass Ausreißer berücksichtigt, der Wert für das 80-Perzentil etwas verringert (20-Perzentil vice versa), d.h. die Unterschiede zwischen beiden Imputationsanwendungen haben etwas größere Auswirkungen auf die Streuungen (Lohnniveaus) zwischen dem 10- und 20-Perzentil sowie dem 80- und 90-Perzentil.

Eine allgemeingültige Antwort auf die Frage, welche Variable nun die „bessere“ ist, gibt es unserer Ansicht nach nicht. Hierauf gehen wir am Schluss dieses Papiers kurz ein und halten an dieser Stelle fest, dass die deskriptiven Vergleiche erste Hinweise geben, dass wohl vor allem an den oberen und unteren Rändern der Lohnverteilung etwas durch die Imputation passiert. Wir gehen möglichen Unterschieden zwischen imputierten und nicht imputierten Werten weiter nach, indem wir Lohnregressionen mit einem konstanten Set an gängigen

[17] Wobei sich die Unterschiede mit je etwa 1,3 Prozent von der Ursprungsverteilung bei einer Ersetzungsquote von etwa 18 Prozent auf einem eher geringen Niveau bewegen.

Kovariaten aufstellen, wo die Einkommensvariablen aus Tabelle 3 die abhängige Variable sind. Anschließend vergleichen wir die entsprechenden Koeffizienten, wobei die Berechnungen jeweils für Männer und Frauen in Ost- und Westdeutschland durchgeführt werden. Um die Darstellung nicht zu überfrachten, diskutieren wir die Verschiebungen der einzelnen Koeffizienten und Standardfehler am Beispiel westdeutscher Männer (vgl. Anhang A3) und geben die Ergebnisse für die anderen Teilgruppen auf Anfrage gerne weiter. Vor dem Hintergrund, dass wir uns insbesondere für den Vergleich der Koeffizienten aufgrund verschiedener Imputationsanwendungen interessieren und demnach die Tabelle A3 eher zeilen- als spaltenweise interpretieren, ist unsere Zusammenfassung der Ergebnisse wie folgt.

Eingegangen wird nur auf signifikante Effekte. In diesen Fällen sinken ausnahmslos die Standardfehler. Zwar verkleinern sich manchmal auch die Koeffizienten (die Bildungsrenditen), aber in der Regel steigen die t-Werte bei der Anwendung der Regressionsgleichung auf die abhängige Variable inklusive der imputierten Werte an[18]. Bei besonders robusten Ergebnissen, die sich in besonders hohen t-Werten ausdrücken – etwa für die Betriebszugehörigkeitsdauer, die Berufserfahrung oder für das Anforderungsniveau der Arbeitsplätze – sind die Ergebnisunterschiede insbesondere der Bildungsrenditen (Koeffizienten), aber auch der Standardfehler, bei einer abhängigen Variable mit und ohne imputierte Werten in aller Regel gering bzw. vernachlässigbar. Es gibt einige wenige und vielleicht inhaltlich weniger „spektakuläre“ Fälle, wo sich ohne Imputation insignifikante Effekte ergeben, mit der Anwendung der beiden Imputationsarten aber das Signifikanzniveau steigt. So ist das (inhaltliche) Ergebnis, dass verwitwete westdeutsche Männer ceteris paribus etwa sieben Prozent mehr verdienen als verheiratete, insignifikant. Das gleiche Ergebnis erhält man, wenn man nur die fehlenden Werte imputiert. Werden aber auch die Ausreißerwerte ersetzt, dann ist der Effekt mit einer etwa fünfprozentigen Irrtumswahrscheinlichkeit signifikant (der t-Wert beträgt in diesem Fall etwa 2,17). Es gibt einige wenige andere Variablen, wo Ähnliches beobachtet werden kann (etwa bei den Freiberuflern in Relation zu Angestellten). Aufgefallen ist uns dabei, dass die Verschiebungen der Signifikanzniveaus insbesondere bei Merkmalen auftreten, die nur wenige Personen haben, gemessen am Gesamtumfang der Stichprobe bzw. dort, wo für die entsprechenden Subsamples angenommen werden kann, dass sie bezüglich ihres Einkommens besonders heterogen strukturiert sind.

Um dies genauer zu untersuchen, wählen wir als Beispiel die Stellung im Beruf. Da die BIBB/BAuA-Erwerbstätigenbefragung 2006 repräsentativ für Erwerbstätige (ohne Auszubildende) ist, befinden sich besonders viele Angestellte in der Stichprobe, d.h. wenn wir die Gleichungen (3.6) und (3.7) auf den Datensatz anwenden, erhalten wir α-Verteilungen, die zu relativ hohen Anteilen auf der Einkommensverteilung der Angestellten

[18] Das ist bei allen Teilgleichungen (und nicht nur bei den westdeutschen Männern) der Fall.

basieren. Diese Verteilung wiederum wird dann auf alle Kategorien der Stellung im Beruf umgelegt und die fehlenden Werte imputiert. Wenn nun gewisse Subgruppen an Erwerbstätigen über besonders heterogene Einkommen verfügen (bzw. besonders häufig die Einkommensangabe verweigern), dann verändert eine Imputation die Verteilung der Einkommen in den jeweiligen Subgruppen stärker, als es bei zahlenmäßig besser vertretenen Subgruppen der Fall ist. Konkret bezogen auf unseren Anwendungsfall stellt sich das für die Stellung im Beruf wie folgt dar (Tabelle 4).

Tabelle 4: Vergleich imputierter und nicht imputierter Werte
(nach Stellung im Beruf, monetäre Angaben in Euro)

	ohne Imputation			Imputation 1			Imputation 2		
Stellung im Beruf	**Fallzahl**	**Mittelwert**	**s.d.***	**Anzahl****	**Mittelwert**	**s.d.***	**Anzahl****	**Mittelwert**	**s.d.***
Arbeiter	4.208	2.020	1.263	514	2.055	1.229	532	2.030	1.079
Angestellte	9.478	2.647	2.008	1.651	2.708	1.918	1.697	2.644	1.538
Beamte	1.473	3.205	2.088	265	3.225	1.946	268	3.151	1.218
selbstständig	1.191	3.576	3.875	420	3.463	3.410	459	3.178	2.060
Freiberufler	360	3.421	3.756	100	3.305	3.395	118	3.057	2.215
freie Mitarbeit	93	1.302	1.185	11	1.383	1.226	12	1.387	1.238
mith. Fam.angehörige	117	964	1.410	74	1.028	1.194	85	1.046	1.178
ZP kann nicht entscheiden	19	1.152	1.300	4	1.306	1.238	4	1.286	1.219
keine Angabe	17	1.872	3.228	5	2.050	3.460	6	2.232	3.537
insgesamt	16.956	2.600	2.154	3.044	2.647	2.045	3.181	2.572	1.542

* Standardabweichung ** *Anzahl* imputierter Werte

Anmerkung: Mittelwerte und Standardabweichungen beziehen sich bei den Imputationen 1 und 2 auf das Gesamtsample
(die Anzahl gültiger Fälle ergibt sich aus der Summe der Spalte Fallzahl plus Spalte Anzahl laut Imputation 1)

Imputation 1: Ersetzung fehlender Werte
Imputation 2: Ersetzung fehlender und Ausreißerwerte

Quelle: BIBB/BAuA-Erwerbstätigenbefragung 2006

Für einzelne Zeilen fallen die Imputationsergebnisse im Vergleich zu den Originalwerten unterschiedlich aus. Erwartungsgemäß sinkt die Streuung der Einkommen, wenn fehlende Werte *und* Ausreißerwerte ersetzt werden. Besonders markant ist der Effekt bei Gruppen, bei denen in Vorhinein eine starke Einkommensstreuung vermutet wurde (Selbstständige, Freiberufler). Bei freier Mitarbeit hingegen ist ein geringer Anstieg der Standardabweichungen zu beobachten.

Wesentlich geringer sind die Effekte auf das Lohnniveau (die Mittelwerte) nach der Anwendung der Imputationsverfahren. Hier sind es insbesondere die Angaben der Selbstständigen und freiberuflich Tätigen, wo angegebene und imputierte Werte gewisse Unterschiede aufweisen, insbesondere, wenn auch Ausreißerwerte ersetzt werden. Bei Arbeitern und Angestellten, bei denen relativ viele Ersetzungen vorgenommen wurden, sind die Abweichungen in der Verteilung zwischen imputierten und nicht imputierten Werten nahezu unbedeutend. Allerdings sollten alle abgebildeten Unterschiede nicht zu hoch

bewertet werden. Insbesondere dort, wo einzelne Auswertungsgruppen ohne imputierte Werte relativ geringe Fallzahlen aufweisen, kann die Arbeit mit den imputierten Werten die Auswertungsqualität verbessern, etwa weil die Analyseergebnisse auf größeren Fallzahlen beruhen, was für inferenzstatistische Schlüsse nicht unbedeutend ist.

4 Diskussion und Ausblick

Unsere Übungen in diesem Papier beschäftigen sich mit der konkreten Anwendung eines in Umfrage- und Prozessdaten häufig vorkommenden Falls des Umgangs mit fehlenden Werten bei einer kontinuierlichen Variable. Neben der Darstellung einfacher Imputationsverfahren haben wir empirisch versucht, den Ausfallmechanismus zu modellieren, der den Daten zugrunde liegt. Wir sind mit dem erhaltenen Ergebnis zufrieden. Die Auswirkungen auf die Lohniveaus halten sich durch die Anwendung des Imputationsverfahrens ausnahmslos in Grenzen. Etwas größer sind die Auswirkungen auf die Verteilung der Einkommensvariable, da wir ein die Verteilung glättendes Verfahren einsetzen und dementsprechend die Standardabweichungen in deskriptiven sowie die Standardfehler in multivariaten Verfahren eine gewisse Reduktion erfahren.

Die Frage ist, ob man durch die Verwendung der imputierten Werte größere Ungenauigkeiten in Kauf nimmt, als wenn man den Ausfallmechanismus der Daten unberücksichtigt lässt. Zumindest im vorliegenden Fall spricht unserer Ansicht nach vieles für die Anwendung des Imputationsverfahrens, weil die Auswirkungen auf die (Verteilung der) abhängigen Variable fast ausnahmslos geringer sind als die Ersetzungsquoten. Konkret betragen die Abweichungen mit und ohne Imputation beim Mittelwert etwa 1,3 Prozent während 18 Prozent aller Beobachtungen ersetzt werden. Das zusätzliche Informationen für knapp 20 Prozent von Gesamt das Analyseergebnis (hier der Auswertung der Einkommensvariable) um etwas über ein Prozent verändern, erscheint plausibel und realistisch, d.h. wir denken, mit unserer Modellierung den tatsächlichen Gegebenheiten ein Stück weit näher gekommen zu sein. Dennoch, wie die Auswertung für die Stellung im Beruf exemplarisch zeigt, kann es für einzelne Subgruppen sinnvoll bleiben, Analyseergebnisse mit und ohne imputierte Werte gegenüberzustellen und diese kritisch gegeneinander abzuwägen.

Die Dokumentation der Imputation einer einzelnen Einkommensvariable erfolgte relativ ausführlich. Zukünftig werden wir uns in den Datenbeschreibungen der BIBB-FDZ-Daten auf die wesentlichen Punkte bei der Anwendung eines Imputationsverfahrens beschränken. Diese verorten wir weniger im methodischen als im empirischen Bereich. Bei Imputationen von kontinuierlichen Merkmalen werden die Veränderungen der Mittelwerte, der Streuung

der entsprechenden Variable und Auswirkungen auf die Signifikanzniveaus bei multivariaten Analysen dokumentiert. Kategoriale Variablen werden wir eher nicht imputieren, weil man sich methodisch auf einem viel schwierigerem Gebiet bewegt[19]. Aber das ist auch oft nicht nötig, weil kategoriale Angaben von Umfrageeinheiten wesentlich seltener verweigert werden als die Angabe eines kontinuierlichen Merkmals.

Literaturverzeichnis

Gabler, Siegfried (2004): Gewichtungsprobleme in der Datenanalyse, Kölner Zeitschrift für Soziologie und Sozialpsychologie 44, 2004, S. 128-147.

Gartner, Hermann (2005): The imputation of wages above the contribution limit of the German IAB employment sample. FDZ-Methodenreport Nr. 2/2005.

Hall, Anja; Tiemann, Michael (2006): BIBB/BAuA-Erwerbstätigenbefragung 2006 – Arbeit und Beruf im Wandel. Erwerb und Verwertung beruflicher Qualifikationen., suf_1.0; Forschungsdatenzentrum im BIBB (Hrsg.); GESIS Köln, Deutschland (Datenzugang); Bundesinstitut für Berufsbildung, Bonn
doi:10.4232/1.4820

Little Roderick J.; Rubin, Donald B. (1987): *Statistical Analysis with Missing Data*. J. Wiley & Sons, New York, 1987.

Molenberghs, G. and Verbeke, G. (2005) Models for Discrete Longitudinal Data. New York: Springer-Verlag.

Rässler, S., Rubin, D.B., Schenker, N. (2008). Incomplete Data: Diagnosis, Imputation, and Estimation, International Handbook of Survey Methodology, de Leeuw, E.D., Hox, J.J., Dillman, D.A. (eds.), Lawrence Erlbaum Associates, Hillsdale, NJ, 370-386.

Rässler, Susanne (2000): Ergänzung fehlender Daten in Umfragen. In: Jahrbücher für Nationalökonomie und Statistik, Bd. 220, H. 1, S. 64-94.

Rubin, Donald B. (1976): Inference and missing data. Biometrica, 63, 581-592.

Verbeke, G. and Molenberghs, G. (2000) Linear Mixed Models for Longitudinal Data. New York: Springer-Verlag.

[19] Etwa weil eine Verteilung des Störterms für einzelne Einheiten konditional anhand von Kovariaten erfolgt. Bei Maximum-Likelihood-Funktionen (mit linearisierten Wahrscheinlichkeiten als abhängiger Variable) ergibt sich ein Problem, weil der Störterm bei unterschiedlichen Ausprägungen auf den unabhängigen Variablen unterschiedliche Verteilungen aufweist.

Anhang

Anhang A1: Ergebnisse einer logistischen Regression, die Einkommensangabe zu verweigern und anzugeben, mehr (gleich eins) oder weniger (gleich null) als 1500 Euro monatlich zu verdienen[20]

	Koeffizient	Standardfehler
Berufliche Stellung (Referenzkategorie: Angestellte/r)		
Arbeiter/in	**0,896**	0,136
Beamter/Beamtin	**-1,314**	0,313
Selbstständige/r	**0,813**	0,177
freiberuflich tätig	0,160	0,316
mithelfende/r Familienangehörige/r	**2,277**	0,489
Keine Angabe	**1,955**	0,885
Wöchentliche Arbeitszeit (Ref.: 35 bis unter 40 h)		
bis 15 h	**2,656**	0,321
15 bis unter 25	**2,300**	0,206
25 bis unter 30	**1,410**	0,272
30 bis unter 35	**1,394**	0,225
40 bis unter 50	-0,234	0,172
ueber 50	**-0,529**	0,202
Migrationshintergrund (Ref.: Deutsche ohne Mig.)		
Deutsche mit Migrationshintergrund	0,228	0,237
Ausländer	-0,372	0,352
Pendeln	-0,479	0,502
Familienstand (Ref.: verheiratet)		
ledig	0,104	0,128
verwitwet	**0,676**	0,326
geschieden	0,313	0,162
Konstante	**-2,056**	0,162

Quelle: BIBB/BAuA-Erwerbstätigenbefragung 2006

Teststatistiken zur Regressionsdiagnostik werden auf Anfrage gerne zur Verfügung gestellt. Signifikante Effekte sind fettgedruckt.

[20] Die Fallzahlen für die Regression entsprechen denen der Tabelle 2 im Haupttext.

Anhang A2: Ergebnisse eines multinominalen Logit-Modells
(baseline: Personen mit gültigen Einkommensangaben; weitere Ausprägungen: Angabe des Einkommens verweigert, Verdienst weniger oder mehr als 1500 Euro im Monat[21])

	Imputation am unteren Rand		Imputation am oberen Rand	
	Koeffizient	Standardfehler	Koeffizient	Standardfehler
Geschlecht	**-0,728**	0,102	**0,102**	0,050
Alter	-0,054	0,049	**-0,136**	0,027
Alter2	0,160	0,120	**0,355**	0,072
Alter3	-0,107	0,093	**-0,270**	0,062
Berufsausbildung (Referenzkategorie: FHS/HS)				
Kein beruflicher Ausbildungsabschluss	**0,578**	0,159	**-0,549**	0,113
Abgeschlossene, auch schulische Berufsausbildung	**0,503**	0,116	**-0,229**	0,052
Berufliche Stellung (Referenzkategorie: Angestellte/r)				
Arbeiter/in	**0,259**	0,106	**-0,487**	0,066
Beamter/Beamtin	**-0,654**	0,285	**-0,210**	0,080
Selbstständige/r	**1,418**	0,125	**0,467**	0,075
freiberuflich tätig	**1,272**	0,193	**0,385**	0,134
mithelfende/r Familienangehörige/r	**2,388**	0,187	0,530	0,276
Keine Angabe	**2,096**	0,591	0,079	0,774
Wöchentliche Arbeitszeit (Ref.: 35 bis unter 40 h)				
bis 15 h	**0,832**	0,183	**-1,579**	0,228
15 bis unter 25	**1,001**	0,156	**-1,098**	0,124
25 bis unter 30	**0,431**	0,211	**-0,528**	0,143
30 bis unter 35	**0,543**	0,182	**-0,387**	0,115
40 bis unter 50	-0,211	0,158	-0,047	0,069
ueber 50	-0,239	0,185	**0,215**	0,078
Betriebszugehörigkeitsdauer	**-0,034**	0,012	**0,038**	0,008
Betriebszugehörigkeitsdauer 2	0,005	0,003	**-0,007**	0,002
Familienstand (Ref.: verheiratet)				
ledig	-0,062	0,113	-0,096	0,056
verwitwet	0,350	0,210	0,049	0,157
geschieden	0,085	0,119	**-0,182**	0,071
Migrationshintergrund (Ref.: Deutsche ohne Mig.)				
Deutsche mit Migrationshintergrund	-0,195	0,179	-0,201	0,105
Ausländer	-0,301	0,242	0,001	0,132
Konstante	**-3,356**	0,712	**-0,428**	0,356

Quelle: BIBB/BAuA-Erwerbstätigenbefragung 2006

Teststatistiken zur Regressionsdiagnostik werden auf Anfrage gerne zur Verfügung gestellt.
Signifikante Effekte sind fettgedruckt.

[21] Die Fallzahlen für die Regression entsprechen denen der Tabelle 2 im Haupttext.

Anhang A3: Vergleich der Koeffizienten und Standardfehler bei Lohnregressionen mit imputierten und nicht imputierten Werten beim Einkommen (westdeutsche Männer[22])

	Koeffizienten			Standardfehler		
	Original	*Imputation 1**	*Imputation 2***	*Original*	*Imputation 1**	*Imputation 2***
Berufliche Stellung (Referenzkategorie: Angestellte/r)						
Arbeiter/in	-0,146	-0,148	-0,134	0,014	0,013	0,011
Beamter/Beamtin	-0,154	-0,151	-0,145	0,019	0,016	0,015
Selbstständige/r	0,234	0,223	0,177	0,042	0,032	0,023
freiberuflich tätig	0,139	0,130	0,228	0,113	0,088	0,039
mithelfende/r Familienangehörige/r	-0,725	-0,645	-0,190	0,365	0,260	0,116
Arbeiter/Angestellter, ZP kann nicht entscheiden	-0,596	-0,422	-0,426	0,245	0,279	0,278
Keine Angabe	0,726	0,688	0,734	0,446	0,338	0,345
Wöchentliche Arbeitszeit (Ref.: 35 bis unter 40 h)						
bis 15 h	-1,285	-1,276	-1,230	0,092	0,086	0,055
15 bis unter 25 h	-0,840	-0,803	-0,799	0,062	0,057	0,044
25 bis unter 30 h	-0,387	-0,365	-0,369	0,075	0,063	0,063
30 bis unter 35 h	-0,252	-0,253	-0,271	0,049	0,044	0,042
40 bis unter 50 h	0,091	0,082	0,081	0,013	0,011	0,010
ueber 50 h	0,211	0,200	0,200	0,017	0,014	0,013
Anforderungsniveau Arbeitsplatz (Ref.: Hochqual. AP)						
Einfach	-0,544	-0,541	-0,526	0,032	0,028	0,026
Qual.	-0,267	-0,261	-0,257	0,022	0,018	0,015
Meister	-0,191	-0,186	-0,166	0,026	0,021	0,016
Migrationshintergrund (Ref.: Deutsche ohne Mig.)						
Deutsche mit Migrationshintergrund	0,017	0,013	-0,017	0,023	0,020	0,017
Ausländer	0,006	0,011	-0,009	0,026	0,023	0,020
Pendeln	-0,175	-0,147	-0,145	0,104	0,093	0,090
Betriebszugehörigkeitsdauer	0,021	0,021	0,020	0,002	0,002	0,002
Betriebszugehörigkeitsdauer 2	-0,004	-0,004	-0,004	0,001	0,001	0,000
Berufserfahrung	0,018	0,018	0,018	0,002	0,002	0,002
Berufserfahrung 2	-0,029	-0,028	-0,031	0,005	0,004	0,004
Familienstand (Ref.: verheiratet)						
ledig	-0,096	-0,094	-0,089	0,012	0,010	0,009
verwitwet	0,067	0,057	0,079	0,050	0,041	0,036
geschieden	-0,151	-0,147	-0,105	0,030	0,026	0,015
Wirtschaftszweig	Ja	Ja	Ja	Ja	Ja	Ja
Beschäftigtenzahl	Ja	Ja	Ja	Ja	Ja	Ja
Berufsausbildung (Ref.: keine berufl. Ausbildung						
Abgeschlossene, auch schulische Berufsausbildung	0,137	0,138	0,142	0,025	0,022	0,020
FHS/HS	0,277	0,276	0,278	0,034	0,029	0,024
Konstante	7,680	7,695	7,683	0,047	0,041	0,034
* Imputation von fehlenden Werten	** Imputation von fehlenden Werten und Ausreißern					

Quelle: BIBB/BAuA-Erwerbstätigenbefragung 2006

Teststatistiken zur Regressionsdiagnostik werden auf Anfrage gerne zur Verfügung gestellt.

[22] Fallzahlen: Original 6.906, Imputation 1 7.809, Imputation 2 7.809

Forschungsdatenzentrum
BIBB.FDZ
im Bundesinstitut für Berufsbildung

Impressum

BIBB-FDZ Daten- und Methodenberichte
Nr. 2/2011
Autoren: Holger Alda, Daniela Rohrbach-Schmidt

Downloads unter:
www.bibb-fdz.de

Herausgeber:
Bundesinstitut für Berufsbildung
Forschungsdatenzentrum
Robert-Schuman-Platz 3
53175 Bonn

Tel.: +49-228-107-2041
Fax: +49-228-107-2020
E-Mail: fdz@bibb.de

Redaktion: Holger Alda
Redaktionsassistenz: Tanja Stiemer

ISSN-Nr.: 2190-300X