Effekte der Multiplen Imputation fehlender Werte am Beispiel von Produktivitätsschätzungen mit dem IAB-Betriebspanel

ARND KÖLLING* & SUSANNE RÄSSLER+

Kurzfassung

Der Beitrag schildert die Effekte von Antwortausfällen bei einzelnen Fragen („item non-response“) auf die Ergebnisse von multivariaten statistischen Analysen. Dabei wird das Verfahren der „Data Augmentation“ angewendet, um die fehlenden Daten zu ersetzen (Schafer 1997). Anhand von Schätzungen der betrieblichen Produktivität wird nun überprüft, ob die „multiple Imputation“ von Beobachtungen (Rubin 1987) zu Veränderungen der Ergebnisse im Vergleich zu den Resultaten mit Antwortausfällen führt. Dazu werden Schätzungen mit Daten des IAB-Betriebspanels aus dem Jahr 2000 durchgeführt. Grundlage der Produktivitätsschätzungen ist eine verallgemeinerte Produktionsfunktion des Translog-Typs (Berndt & Christensen 1973), wobei Arbeit und Kapital als Produktionsfaktoren unterstellt werden. Zusätzlich werden Brancheneinflüsse und die Effekte der Nutzung moderner Technologien berücksichtigt. Von besonderem Interesse bei der Untersuchung sind Differenzen zwischen den alten und den neuen Bundesländern. Um zu überprüfen, ob nicht nur ein konstanter Unterschied zwischen Ost und West existiert, werden neben einer Dummy-Variablen auch alle anderen exogenen Variablen mit dieser Ost-West-Variable multipliziert und gehen als Interaktionsvariablen in die Schätzung ein. Die Funktion wird als Maximum-Likelihood-Funktion geschätzt, in der ein multiplikatives Modell der Heteroskedastie unterstellt wird (Greene 1998). Die Ergebnisse bestätigen das zugrunde liegende Modell. Neben einer Reihe von Unterschieden in den Branchen, ergibt sich auch eine höhere Produktivität bei Nutzung von modernen Technologien. Zwischen Ost und West zeigen sich dagegen nur Unterschiede in der Konstanten und bei einigen Branchen, nicht jedoch bei den Produktionsfaktoren. Durch den Antwortausfall scheinen diese Differenzen zu hoch ausgewiesen zu werden. Bei den ergänzten Daten reduzieren sich die konstanten Unterschiede in der Produktivität um etwa 11%-Punkte. Ebenso werden die Branchendifferenzen deutlich geringer.

Key words: Data Augmentation, multiple Ergänzung, Markov Chain Monte Carlo, Translog-Produktionsfunktion.

JEL-Klassifikation: C15, C81, D24

* Institut für Arbeitsmarkt- und Berufsforschung (IAB), Regensburger Str. 104, 90478 Nürnberg, ✆ 0911 / 179-3174, mail: Arnd.Koelling@iab.de.

+ Universität Erlangen-Nürnberg, Lehrstuhl für Statistik und Ökonometrie, Lange Gasse 20, 90403 Nürnberg, ✆ 09 11 / 53 02-276, mail: Susanne.Raessler@wiso.uni-erlangen.de

1 Einleitung

Ein typisches Problem der multivariaten Analyse auf Basis von Umfragedaten sind die unterschiedlichen Lücken in einem solchen Datensatz, wie sie auch im IAB-Betriebspanel auftreten. In diesem Beitrag wird eine Regression geschätzt werden, die die ausfallbehafteten Variablen wie Umsatz, den prozentualen Anteil der Vorleistungen oder Erweiterungsinvestitionen einer Unternehmung enthält. Die gängigen Statistik-Softwareprogramme eliminieren zur Berechnung der Schätzfunktionen jede Zeile, d.h. Beobachtung, in der mindestens ein Wert fehlt. Zur Schätzung werden also nur diejenigen „verfügbaren" Beobachtungen verwendet, die bei allen in der Schätzung verwendeten Variablen vollständige Angaben enthalten. Im Falle der Schätzung einer Translog-Produktionsfunktion mit den Daten der 8. Welle des IAB-Betriebspanels reduziert sich daher das verfügbare Datenmaterial der Nettosstichprobe auf 59%, nämlich genau 6489 Beobachtungen. Auf Grund der Fallreduktion ist zu fragen, ob diese 59% noch „repräsentativ" sind für die Grundgesamtheit und ob nicht wertvolle Informationen verworfen werden. Zur Erörterung dieser Frage empfiehlt es sich zunächst, eine Strukturierung des Datenausfalls vorzunehmen.

In Anlehnung an Rubin (1987) und Little und Rubin (1987) lassen sich folgende Fälle der Antwortverweigerung klassifizieren:

- MCAR (missing completely at random),
- MAR (missing at random),
- MNAR (missing not at random).

MCAR beschreibt einen rein zufälligen Antwortausfall, etwa wenn der Eintrag des Interviewten bei der Dateneingabe nicht lesbar ist, vorausgesetzt, der Interviewte hat nicht mit Absicht „geschmiert". Bei MAR kann der Datenausfall zumindest durch andere beobachtete Variablen „erklärt" werden. Dies ist beispielsweise der Fall, wenn die Umsatzangabe bei Großbetrieben häufiger verweigert wird als bei Kleinbetrieben, und die Betriebsgröße etwa gemessen durch die Beschäftigenzahl bekannt ist. Analytisch ist der MNAR-Ausfall am Schwierigsten zu behandeln. In diesem Fall würden Betriebe den Umsatz beispielsweise dann mit einer höheren Wahrscheinlichkeit nicht angeben, wenn sie einen besonderes hohen Umsatz haben. Ferner wird meist unterstellt, daß sich die Parameter ξ des den Datenausfall generierenden Prozesses und die Parameter θ des Modells der vollständigen Daten nicht gegenseitig beeinflussen, d.h. variationsfrei sind. Liegen diese Variationsfreiheit und mindestens ein MAR-Ausfall vor, spricht man von einem „ignorierbaren Datenausfallmechanismus". Ignorierbar deshalb, weil man nun zeigen kann, daß Likelihood-basierte Schlußfolgerungen über den

interessierenden Parameter θ des Modells der vollständigen Daten möglich sind, ohne ein Modell für den Antwortmechanismus explizit spezifizieren zu müssen.

Genauere Analysen des Datenausfalls im IAB-Betriebspanel legen die Vermutung eines MAR-Ausfalles nahe, daher soll ein ignorierbarer MAR-Datenausfall unterstellt werden. Es ist zu beachten, daß es nicht möglich ist, die Annahme der Ignorierbarkeit an den beobachteten Daten zu überprüfen. Die MAR-Annahme entspricht also der „schwächsten“ Annahme über den Datenausfall, wenn der Ausfallmechanismus nicht explizit modelliert werden soll oder kann. Eine Schätzung auf Basis nur der verfügbaren Fälle trifft hingegen die „strengste“ Annahme, den MCAR-Ausfall. Ferner gehen u.U. wertvolle Informationen verloren, da Einheiten mit fehlenden Werten aus der Schätzung ganz entfernt werden. Deshalb ist es naheliegend, anstelle alle, d.h. auch die beobachteten Werte solcher lückenhaften Objekte zu entfernen, lieber die Lücken aufzufüllen, d.h. zu ergänzen. Im englisch sprachigen Raum hat sich dafür der Begriff der „Imputation“ eingebürgert, der sich ebenso im Deutschen verwenden läßt. Es gibt nun eine Vielzahl von Ergänzungstechniken, die alle ihre Vor- und auch Nachteile haben, vgl. Rässler (2000). Rubin (1987) führt sein Verfahren der „Multiplen Imputation“ (kurz: MI) ein. MI hat, im Gegensatz zu den üblichen ad hoc Techniken, eine theoretische Fundierung und sich in vielen Evaluationen und Untersuchungen immer wieder als überlegen erwiesen, vgl. beispielweise Schafer (1997). Mangelnde Verfügbarkeit entsprechender Software und die durchaus anspruchsvolle Theorie mögen dafür verantwortlich sein, daß sich die Technik des MI noch nicht allgemein in der Praxis durchgesetzt hat.

In dieser Untersuchung wird erstmals die Technik der multiplen Ergänzung auf deutsche Betriebsdaten angewandt. Dies erlaubt es einerseits, bei der Schätzung alle verfügbaren Information zu verwenden sowie zusätzlich andererseits die Unsicherheit des Datenausfalls und damit der Ergänzung widerzuspiegeln. Der Artikel ist wie folgt strukturiert. Im zweiten Abschnitt wird zunächst die Datenlage und der Datenausfall beschrieben. Anschließend folgt die Darlegung des Prinzips der Multiplen Ergänzung. Nach der Diskussion geeigneter Ergänzungsverfahren werden im vierten Abschnitt das Analysemodell sowie das Ergänzungsmodell diskutiert und die Erzeugung der Ergebnisse unter Verwendung der MI-Schätzer erklärt. Abschnitt fünf präsentiert die Ergebnisse der Schätzung auf Basis nur der verfügbaren Daten und vergleicht diese mit den Ergebnissen auf Basis der ergänzten Daten. Eine Zusammenfassung gibt schließlich Abschnitt sechs.

2 Das Antwortverhalten im IAB-Betriebspanel

Als Grundlage der Untersuchung stehen Angaben aus dem IAB-Betriebspanel zur Verfügung. Der Datensatz wird seit 1993 im Auftrag des Instituts für Arbeitsmarkt- und Berufsforschung der Bundesanstalt für Arbeit (IAB) erhoben (vgl. Kölling 2000). Die Grundgesamtheit

der Stichprobenziehung besteht aus allen Betrieben mit mindestens einem sozialversicherungspflichtigen Beschäftigten. Aus der Grundgesamtheit ausgeschlossen sind daher Betriebe ohne sozialversicherungspflichtige Beschäftigte, z.B. sogenannte Scheinselbständige, Betriebe allein mit Beschäftigten selbständiger Versicherungsarten (Bergleute, Landwirte, Künstler, Publizisten) oder Dienststellen im öffentlichen Sektor, in denen ausschließlich Beamte beschäftigt sind. Das Sample wird als eine geschichtete Stichprobe aus der Betriebsdatei der Bundesanstalt für Arbeit gezogen. In dieser Datei werden die gesetzlichen Pflichtmeldungen der Arbeitgeber an die Sozialversicherungsträger über eine Betriebsnummer zu örtlichen “Betriebseinheiten” aggregiert und können hinsichtlich verschiedener Merkmale wie Wirtschaftszweigzugehörigkeit und Betriebsgröße abgegrenzt werden. Bei der Stichprobenziehung wird das Verfahren der ‘optimalen Schichtung‘ (varianzminimierend) auf eine Matrix aus 20 Branchen und 10 Betriebsgrößenklassen angewendet. Dadurch sind Großbetriebe überproportional im Datensatz vertreten. Jedoch besteht der Datensatz weiterhin zum größten Teil aus Kleinbetrieben:

Tabelle 1: Auswahlwahrscheinlichkeiten im IAB-Betriebspanel (1993)

Anzahl der sozialversicherungspflichtigen Beschäftigten	Bruttostichprobe	Auswahlwahrscheinlichkeiten	Nettostichprobe	Antwortquoten
1 – 4	927	0,0011	625	0,67
5 – 9	390	0,0015	250	0,64
10 – 19	423	0,0030	299	0,71
20 – 49	777	0,0089	542	0,70
50 – 99	486	0,0153	350	0,72
100 – 199	491	0,0304	376	0,77
200 – 499	829	0,0862	615	0,74
500 – 999	426	0,1504	304	0,71
1000 – 4999	1286	0,8765	924	0,72
> 5000	97	0,9127	71	0,73
Total	6132	0,0043	4356	0,71

Quelle: IAB-Betriebspanel 1993

Um Neugründungen abzubilden, werden dem Panel jedes Jahr Betriebe hinzugefügt, die zum ersten Mal einen sozialversicherungspflichtigen Beschäftigten eingestellt haben. In regelmäßigen Abständen wird zusätzlich eine Ergänzungsstichprobe gezogen, um die Folgen der Panelmortalität auszugleichen. Das Panel ist weitgehend regionalisiert, d. h. für die meisten Bundesländer sind detaillierte Analysen möglich. Von 1993 bis 1995 wurde das IAB-Betriebspanel ausschließlich in den alten Bundesländern erhoben. Seit 1996 werden auch die neuen Bundesländer einbezogen, für die seitdem regionale Auswertungen auf Bundeslandsebene möglich sind. Für 2000 können diese kleinräumigeren Analysen auch für eine

Reihe von westdeutschen Bundesländern durchgeführt werden. Es wird angestrebt, für alle Bundesländer repräsentative Stichproben bereitzustellen. Die Befragung wird überwiegend als persönliches Interview durchgeführt. In Hamburg und Bremen wurden eine Reihe von Betrieben postalisch befragt, da aufgrund der städtischen Struktur eine vergleichsweise große Stichprobe notwendig ist, um repräsentative Ergebnisse zu erhalten. Zusätzlich waren erstmals 405 Betriebe des Produzierenden Gewerbes aus Niedersachsen Bestandteil des IAB-Betriebspanels, die bis 1997 zur Stichprobe des Hannoveraner Firmenpanels gehörten. In der 8. Welle (2000) konnten so annähernd 14.000 Betriebe befragt werden. Dies entspricht ungefähr 0,6% der Betriebe mit 9,5% der sozialversicherungspflichtigen Beschäftigten in Deutschland.

Die Teilnahmebereitschaft der angesprochenen Betriebe liegt bei rund 60%. Diese Zahl wird jedoch stark durch die postalischen Interviews beeinflußt. Erwartungsgemäß antwortete in 2000 nur jeder vierte Betrieb, wenn kein persönliches Interview geführt wurde. Beschränkt man sich nur auf die persönlichen Interviews ergibt sich eine Antwortrate von annähernd 70%. Dieser Wert liegt im üblichen Rahmen der einzelnen Erhebungswellen des IAB-Betriebspanels, wobei sich allerdings Differenzen zwischen den neuen und den alten Bundesländern ergeben. Im Osten liegt die Antwortbereitschaft bei über drei Viertel der Betriebe und damit deutlich über der im Westen. Dies lässt sich aus dem recht hohen Anteil an neu befragten Betrieben erklären. Mehr als 11.500 Betriebe in den alten Bundesländern wurden erstmals angesprochen. Dies sind zehnmal so viele Betriebe wie in den neuen Bundesländern. Allerdings zeigen sich auch hier Ost/West-Differenzen. Bei den persönlichen Interviews ist die Antwortbereitschaft im Osten um ca. 11%-Punkte höher. Die in einigen Bereichen Westdeutschlands durchgeführte postalische Befragung wird jedoch nur von rund 25% der Betriebe beantwortet. Dies liegt allerdings im Rahmen der üblichen Rücklaufquoten für diese Erhebungsform. Bei den Einheiten, die bis 1997 im Rahmen des Hannoveraner Firmenpanels befragt wurden, lag die Teilnahmebereitschaft bei 57% und damit über dem westdeutschen Durchschnitt. Üblicherweise ist in Panelerhebungen das Antwortverhalten von Betrieben, die bereits an der Befragung teilgenommen haben, sehr viel besser als bei der Erstbefragung. So auch beim IAB-Betriebspanel. In jeder Panelwelle lag die Quote der Beantwortung von wiederholt befragten Betrieben bei über 80%. In 2000 war sie sogar größer als 84%. Dabei sind die Unterschiede zwischen Ost und West gering. Die sogenannte Nachbearbeitungsstichprobe stellt nur eine kleine Gruppe dar. Hierbei handelt es sich um Betriebe, die im Vorjahr nicht an der Befragung teilgenommen haben, aber Bereitschaft signalisiert haben, dies im Jahr 2000 wieder zu tun. Erwartungsgemäß verbergen sich in dieser Gruppe viele Betriebe, die kein Interesse mehr an der Befragung haben. Untersuchungen des Antwortausfalls mit den Paneldaten haben gezeigt, daß nur wenige Faktoren eine Rolle spielen (Hartmann & Kohaut 2000). Insbesondere der Wechsel des Interviewers kann zu einer Ver-

weigerung der Betriebe führen. Es wird deutlich, daß konstante Strukturen der Befragung notwendig sind, um die Teilnahmebereitschaft der Betriebe zu erhalten.

Tabelle 2: Bruttostichprobe und auswertbare Interviews - West / Ost

Teilstichprobe	Gebiet	Brutto abs.	Auswertbare Interviews abs.	% vom Brutto
Antworter aus Welle 7	West	4.510	3.744	83,0%
	Ost	5.405	4.593	85,0%
	Gesamt	**9.915**	**8.337**	**84,1%**
Nachbearbeitungsstichprobe 2000	West	431	142	32,9%
	Ost	271	101	37,3%
	Gesamt	**702**	**243**	**34,6%**
Ergänzungsstichprobe 2000	West ohne Mail	6.387	3.219	50,4%
	West nur Mail	5.381	1.388	25,8%
	Ost	1.467	896	61,1%
	Gesamt	**13.235**	**5.503**	**41,6%**
	Gesamt ohne Mail	**7.854**	**4.115**	**52,4%**
Gesamt	West	16.709	8.493	50,8%
	Ost	7.143	5.590	78,3%
	Gesamt	**23.852**	**14.083**	**59,0%**
	Gesamt ohne Mail	**18.471**	**12.695**	**68,7%**

Quelle: Infratest Burke Sozialforschung

Im Gegensatz zum Ausfall ganzer Einheiten, sind manche Betriebe nicht bereit, alle Fragen zu beantworten. Dieser „Item-nonresponse“ kann einerseits durch das Nichtwissen oder die Überforderung der Befragten hervorgerufen werden, andererseits berühren manche Fragen möglicherweise sensible Bereiche wie die Lohn- oder Investitionssumme. Die folgende Tabelle 3 enthält die Angaben mit den häufigsten Antwortausfällen.

Die häufigsten Antwortausfälle ergeben sich bei dem Anteil der Vorleistungen am Umsatz. Diese Variable verzeichnet zusammen mit den im Jahr 2000 nicht abgefragten bezahlten Überstunden, traditionell den höchsten Anteil fehlender Angaben. Um dem zu begegnen wurde die Frage in früheren Wellen zunehmend präzisiert. Es scheint jedoch, daß insbesondere kleinere und traditionell strukturierte Betriebe durch die Frage überfordert sind.

Tabelle 3: Fragen mit hohen Antwortausfällen (KA)

Frage / Variable	Inhalt	Einheit	KA-Anteil *) 2000	1999
H17	Anteil Vorleistungen am Umsatz	(%)	34 %	29 %
H22a-d	Regionale Herkunft der Investitionsgüter[1]	(Kategorien)	5-17 %	-
H12	Geschäftsvolumen	(DM)	13 %	13 %
H41	Summe der Investitionszuschüsse	(DM)	12 %	12 %
H20	Anteil Erweiterungsinvestitionen	(%)	10 %	11 %
H45	Lohn- und Gehaltssumme Juni	(DM)	10 %	11 %
H36	Gründe warum der Betrieb nicht ausbildet	(Antwortvorg.)	10 %	-
H59c	Reaktion auf unbesetzte Stellen	(Antwortvorg.)	9 %	-
H63b	Dem Arbeitsamt gemeldete Stellen	(Anzahl)	9 %	-
H65b	Frauenanteil bei Personalabgängen	(Anzahl)	7 %	-
H13proz	Entwicklung Geschäftsvolumen ggü. Vorjahr	(%)	7 %	7 %

Quelle: Infratest Burke Sozialforschung. *) In % der Fälle (ungewichtet), die die entsprechende Frage zu beantworten hatten.

Da im weiteren Verlauf der Studie die betriebliche Produktivität gemessen werden soll, wird diese Angabe jedoch benötigt, um die Wertschöpfung des Betriebes zu bestimmen. Daher eignet sie sich besonders, um die Effekte der Multiplen Imputation darzustellen. Weitere Variablen, die bei den Produktivitätsschätzungen verwendet werden und viele fehlende Werte aufweisen, sind das Geschäftsvolumen (Umsatz), der Anteil der Erweiterungsinvestitionen und die Lohn- und Gehaltssumme. Bei diesen Fragen kann man annehmen, daß zumindest zum Teil die Bereitschaft fehlt, eine Angabe zu machen. Auch hier sollten daher die Effekte der Datenergänzung betrachtet werden.

3 Prinzip der multiplen Ergänzung

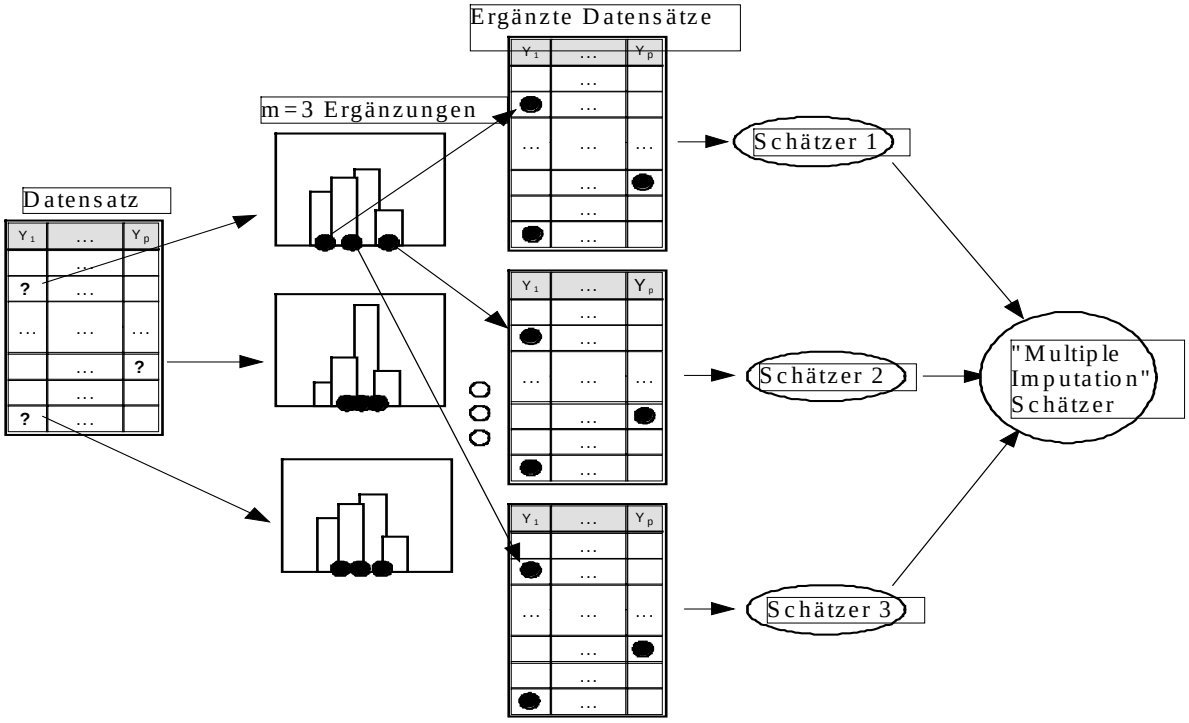
3.1 Grundidee

Die Idee der multiplen Ergänzung ist so einfach wie ansprechend. Bei einer einzigen Ergänzung, die zu einem vervollständigten Datensatz führt, wird dieser üblicherweise so analysiert, als wären die Daten bereits vollständig gewesen. Das dem Datenausfall immanente Identifikationsproblem wie es Manski (1995) diskutiert, wird also völlig ignoriert. Bei der multiplen Ergänzung hingegen werden "plausible" Werte für die fehlenden Daten mehrfach ergänzt (imputiert). Dadurch wird nicht nur ein vervollständigter Datensatz erzeugt sondern mehrere. Diese Datensätze können jeweils einzeln mit den üblichen Verfahren ausgewertet werden,

[1] Besonders hoch sind die Antwortausfälle bei dieser Frage bei den schriftlich befragten Betrieben. Offenbar war vielen Befragten die Logik der Frage unklar, so daß fehlende Angaben hier in der Regel als „gar nicht“ zu interpretieren ist.

wobei die dann notwendigerweise schwankenden Schätzwerte die Unsicherheit des Datenausfalls und damit der Ergänzung widerspiegeln. Die verschiedenen Ergebnisse lassen sich schließlich nach der von Rubin (1987) entwickelten Theorie kombinieren. Die Grundidee der multiplen Ergänzung läßt sich am besten graphisch veranschaulichen, siehe Abbildung 1. Eine ausführliche Einführung in die multiple Ergänzung findet der interessierte Leser insbesondere bei Schafer (1997, 1999a) und Brand (1999).

Abbildung 1: Grundidee der multiplen Ergänzung



Die Erzeugung der verschiedenen, zu ergänzenden Werte kann auf vielfältige Weisen erfolgen. Rubin (1987) diskutiert eine Reihe von Möglichkeiten, Schafer (1997) formuliert mit dem Verfahren der „Data Augmentation“ sehr flexible Ergänzungstechniken, die hier verwendet werden sollen. Die Kombination der Ergebnisse ist prinzipiell einfach und in Abbildung 2 illustriert. Es sei θ ein Parameter der Grundgesamtheit, wie etwa der Mittelwert oder ein Regressionsparameter. Ferner sei $\hat{\theta}$ ein Schätzer für θ, der bei vollständigen Daten verwendet worden wäre. Wichtig ist dabei, daß es sich bei den jeweiligen Schätzern $\hat{\theta}$ um Schätzfunktionen handelt, für die zumindest approximativ die Normalverteilung gilt, d.h.

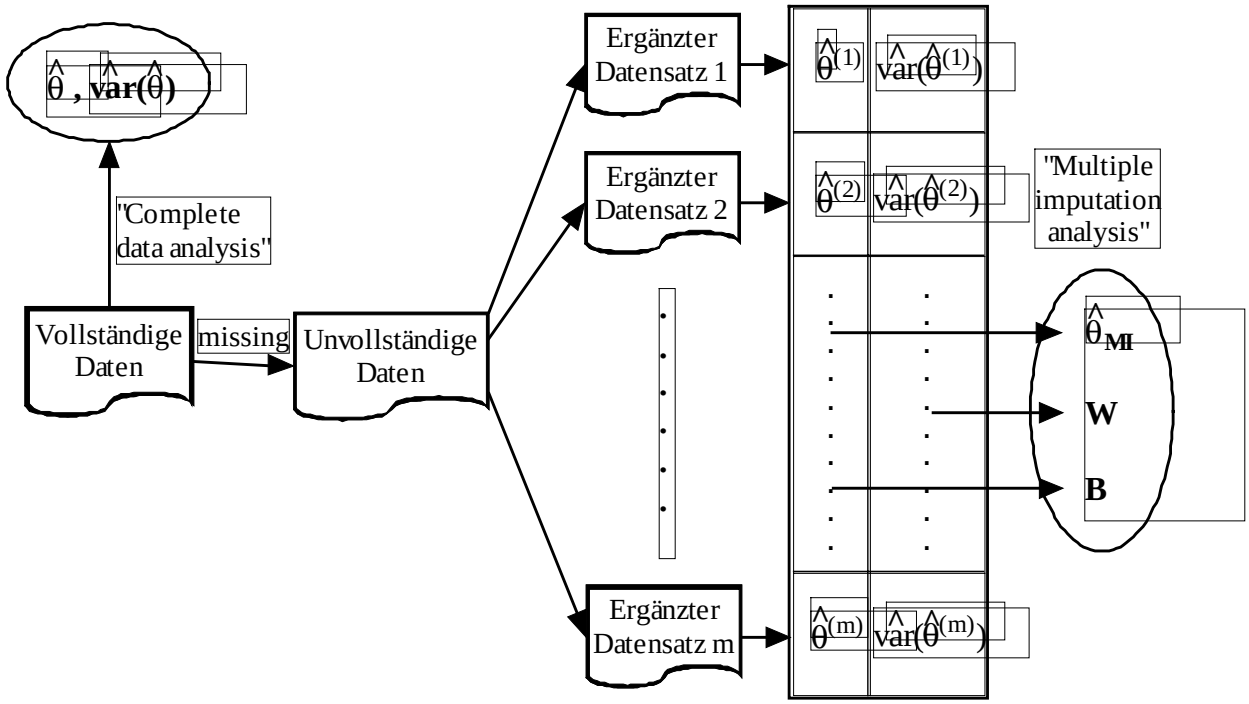
$$(1) \qquad (\hat{\theta} - \theta) \Big/ \sqrt{\hat{V}(\hat{\theta})} \sim N(0,1) .$$

Ggf. sollte die Schätzfunktion geeignet transformiert werden, um diese asymptotische Normalverteiltheit zu gewährleisten. Wurden m Datensätze ergänzt, können daraus m Schätzer

$\hat{\theta}^{(i)}$ und ihre üblichen Varianzschätzer $\hat{V}(\hat{\theta}^{(i)})$ berechnet werden. Der MI-Schätzer $\hat{\theta}_{MI}$ für einen Parameter θ ergibt sich nun aus dem einfachen ungewichteten Mittel der einzelnen Schätzergebnisse. Um einen Standardfehler für $\hat{\theta}_{MI}$ zu erhalten, wird die „totale" Varianz T des MI-Schätzers nach dem Prinzip der Streuungszerlegung ermittelt, wie folgt:

- W = "Within Imputation" Varianz, mit $W = \frac{1}{m}\sum_{i=1}^{m}(\hat{\theta}^{(i)} - \hat{\theta}_{MI})^2$
- B = "Between Imputation" Varianz, mit $B = \frac{1}{m-1}\sum_{i=1}^{m}\hat{V}(\hat{\theta}^{(i)})$
- T = "Gesamte" Varianz = $W + (1+\frac{1}{m})B$.

Abbildung 2: Das Prinzip der multiplen Ergänzung



Hypothesentests und Konfidenzintervalle können dann auf Basis einer t-Verteilung ermittelt werden, mit

(2) $$(\hat{\theta}_{MI} - \theta)/\sqrt{T} \sim t_v$$

und Freiheitsgraden $v = (m-1)\left(1+\frac{W}{(1+m^{-1})B}\right)^2$.

Für mehrdimensionale Tests und Konfidenzintervalle wird die Theorie der multiplen Ergänzung etwas komplexer. Erste Verfahren werden beispielsweise von Li et al. (1991) sowie von Meng und Rubin (1992) vorgeschlagen, einen Überblick gibt auch Schafer (1997, 112ff.).

Der große Vorteil der multiplen Ergänzung liegt zunächst in der einfachen Analyse und der Verwendung von sämtlicher zur Verfügung stehender Information. Die Daten können mit jeder, für die vollständigen Daten geeigneten Methode analysiert werden. Ob diese Analyse allerdings zu, im frequentistischen Sinne validen Aussagen führt, hängt von der gewählten Ergänzungstechnik ab. Rubin (1987, 118f.) definiert dazu den Begriff der „proper multiple-imputation methods“, der im folgenden als „frequentistisch geeignete“ MI-Methoden übersetzt werden soll. Grob gesprochen führt eine frequentistisch geeignete MI-Methode dazu, daß unter wiederholten Ergänzungen dieselben Schlußfolgerungen zu erwarten sind, wie man sie erhalten würde, wären die Daten vollständig gewesen. Nun sind der Beweis einer Ergänzungstechnik frequentistisch geeignet zu sein und insbesondere auch die konstruktive Herleitung einer solchen Methode sehr schwierig. Daher schlägt Rubin (1987, 125ff.) vor, sich der Bayes' Statistik zu bedienen.

3.2 Geeignete Ergänzungsverfahren

Es bezeichne Y_{obs} die beobachteten Variablen und Y_{mis} die nicht beobachteten Variablen jeder Einheit i einer Stichprobe von i= 1,2,..., n, vgl. Tabelle 4.

Tabelle 4: Beispiel einer Beobachtung mit fehlenden Werten

Nr.	West/Ost	Branche	Umsatz (in Mio)	% Anteil Vorleistung	Beschäftigung	Erweit.-Investition	...
...	...	...	...	...	...	...	...
i	obs	obs	mis	mis	obs	obs	...
...	...	...	...	...	...	...	...

Die konstruktive Anleitung zur Ermittlung eines geeigneten Ergänzungsverfahrens läßt sich wie folgt grob skizzieren:

1. Formuliere den datenerzeugenden Prozeß für die vollständigen Daten Y mit $f(y|\theta)$ und entsprechend für die beobachteten Daten Y_{obs} mit $f(y_{obs}|\theta)$.

2. Verwende eine geeignete a priori Verteilung $f(\theta)$ für die unbekannten Parameter θ (die jetzt als Zufallsvariable betrachtet werden) des Datenmodells und berechne gemäß des

Theorems von Bayes die „complete-data“ a posteriori Verteilung $f(\theta|y)$ und die „observed-data“ a posteriori Verteilung $f(\theta|y_{obs})$.

3. Erzeuge m Zufallszüge aus der prädiktiven a posteriori Verteilung $f(y_{mis}|y_{obs})$ der fehlenden Daten Y_{mis} gegeben die beobachteten Daten y_{obs}.

Schafer (1997, 105) definiert den Begriff der "Bayesianly proper" (bayesianisch geeignete) MI Methoden, wenn ihnen m unabhängige Zufallszüge für die fehlenden Daten aus ihren prädiktiven a posteriori Verteilungen

(3) $$f(y_{mis}|y_{obs}) = \int f(y_{mis}|y_{obs},\theta) f(\theta|y_{obs}) d\theta$$

zugrunde liegen. Diese können beispielsweise direkt realisiert werden,

1. durch Zufallszüge für θ aus den observed-data a posteriori Verteilungen $f(\theta|y_{obs}) = L(\theta;y_{obs}) f(\theta)/f(y_{obs})$, wobei $f(y_{obs})$ als Normierungskonstante bezeichnet wird, und

2. durch Zufallszüge für Y_{mis} aus den prädiktiven bedingten Verteilungen $f(y_{mis}|y_{obs},\theta)$ für aktuelle Werte von θ (mit (1) generiert).

Das Problem bei diesem Vorgehen liegt üblicherweise in der Komplexität von $f(\theta|y_{obs})$. Diese observed-data a posteriori Verteilungen sind meist sehr unhandlich und schwierig zu berechnen. Oftmals ist es wesentlich einfacher, die complete-data a posteriori Verteilungen zu bestimmen als für jedes Ausfallmuster die observed-data a posteriori Verteilungen anzugeben. Daher werden insbesondere von Schafer (1997) in seinem Buch iterative Lösungen über einen Spezialfall des Gibbs-Samplers, genauer den Data Augmentation Algorithmus vorgeschlagen. Der Terminus Technicus Data Augmentation geht auf den grundlegenden Artikel von Tanner und Wong (1987) zurück. Sie entwickeln diesen Algorithmus, um bei unvollständigen Daten Zufallszüge der Parameter aus ihren a posteriori Verteilungen generieren zu können.

Der Data Augmentation Algorithmus läßt sich unter die Markov Chain Monte Carlo Methoden einordnen sowie als Bayes’sche stochastische Variante des bekannten EM-Algorithmus‘ darstellen. Analog zum EM-Algorithmus werden iterativ immer wieder zwei Schritte, der **I**mputations- und der **P**osterior-Schritt, wie folgt durchlaufen:

I-Schritt: Generiere Zufallszüge der fehlenden Daten aus ihren prädiktiven bedingten Verteilungen $f(y_{mis}|y_{obs},\theta^{(t)})$ für aktuelle Werte von $\theta^{(t)}$.

P-Schritt: Erzeuge Zufallszüge der Parameter aus ihren complete-data a posteriori Verteilungen $f(\theta|y_{obs},y_{mis}^{(t)})$ für aktuelle Werte von $y_{mis}^{(t)}$.

Nach einem Einschwingvorgang und unter schwachen Regularitätsbedingungen sollte sich die Unabhängigkeit der Kette von den Startwerten sowie die Konvergenz der Kette mit den stationären Verteilungen $f(\theta|y_{obs})$ und $f(y_{mis}|y_{obs})$ bei $t \to \infty$ einstellen, für eine ausführliche Diskussion siehe etwa Schafer (1997, 72ff.) oder van Dyk und Meng (2001). Eine konkrete Modellierung wird im folgenden Abschnitt anhand der IAB-Daten gezeigt.

4 Data Augmentation im IAB-Betriebspanel

4.1 Das Ergänzungsmodell

Zur Erzeugung der Ergänzungen wird die stand-alone Software NORM 2.03 von Schafer (1999b) verwendet. Es wird eine r-dimensionale Normalverteilung als Datenmodell unterstellt bei n unabhängigen Beobachtungen und r beobachtbaren Variablen, d.h. für die beobachtbare r-dimensionale Variable Y_i jeder Einheit i des Datensatzes gilt

(4) $$Y_i \sim N(\mu,\Sigma),$$

für i=1,2,...n und stochastisch unabhängig. Als a priori Verteilung $f(\mu,\Sigma)$ für die Parameter μ und Σ kommen die üblichen näherungsweise unabhängigen und uninformativen a priori Verteilungen zur Anwendung mit

(5) $$f(\mu,\Sigma) \approx f(\mu)f(\Sigma) \propto |\Sigma|^{-(r+1)/2}.$$

Als a posteriori Verteilungen $f(\mu,\Sigma|y)$ der Parameter gegeben die Daten stellen sich dann eine Normalverteilung für μ gegeben Σ und eine inverse Wishard-Verteilung für Σ ein, mit

(6) $$\begin{aligned}\Sigma|y &\sim W^{-1}(n-1,(nS(\bar{y}))^{-1}),\\ \mu|\Sigma,y &\sim N(\bar{y},\Sigma/n),\end{aligned}$$

wobei $S(\bar{y}) = \frac{1}{n}\sum_{i=1}^{n}(y_i+\bar{y})(y_i-\bar{y})'$ die Stichprobenkovarianzmatrix beschreibt mit $\bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_i$ und $y_i = (y_{i1},y_{i2},...,y_{ir})'$. Die bedingten prädiktiven Verteilungen der fehlenden

Daten gegeben die beobachteten Daten ergeben sich hier als bedingte Normalverteilungen gemäß

(7) $$Y_{mis}\left|y_{obs},\mu,\Sigma \sim N(\mu_{mis|obs},\Sigma_{mis|obs}\right.)$$

Im Data Augmentation Algorithmus werden die folgende zwei Schritte iterativ durchlaufen:

I-Schritt: Für jede Einheit i, die fehlenden Werte aufweist, wird für die fehlenden Werte ein Zug aus der bedingten prädiktiven Verteilung $f(y_{mis}\left|y_{obs},\theta\right.)$ gezogen analog zu (7) für aktuelle Parameterwerte $\mu^{(t)}$ und $\Sigma^{(t)}$ nach

(8) $$Y_{mis}{}^{(t)}\left|y_{obs},\mu^{(t)},\Sigma^{(t)} \sim N(\mu_{mis|obs}{}^{(t)},\Sigma_{mis|obs}{}^{(t)}\right.)$$

P-Schritt: Auf Basis dieser vervollständigten Daten $y^{(t)} = (y_{obs}, y_{mis}{}^{(t)})$ können nun aktuelle Werte für den Vektor $\bar{y}^{(t)}$ der Stichprobenmittelwerte sowie für die Stichprobenkovarianzmatrix $S(\bar{y}^{(t)}) = \frac{1}{n}\sum_{i=1}^{n}(y_i^{(t)} + \bar{y}^{(t)})(y_i^{(t)} - \bar{y}^{(t)})'$ berechnet werden. Dann sind Zufallszüge der Parameter zu generieren aus ihren complete-data a posteriori Verteilungen analog zu (6)

(9)
$$\Sigma^{(t+1)}\left|y^{(t)} \sim W^{-1}(n-1,(nS(\bar{y}^{(t)}))^{-1}\right.),$$
$$\mu^{(t+1)}\left|\Sigma^{(t+1)},y^{(t)} \sim N(\bar{y}^{(t)},\Sigma^{(t+1)}/n\right.).$$

Die ausführliche Herleitung dieser Verteilungen kann bei Schafer (1997, 148ff.) oder auch bei Rässler (2001, 96ff.) nachgelesen werden. Die beiden Schritte (8) und (9) werden nun ausgehend von Startwerten $\mu^{(0)}$ und $\Sigma^{(0)}$ iterativ so lange durchlaufen, bis sich Unabhängigkeit von den Startwerten einstellt und die Konvergenz der Markov Kette angenommen werden kann. Für $t \to \infty$ werden als stationäre Verteilungen gerade die gewünschten Verteilungen $f(\theta\left|y_{obs}\right.)$ und $f(y_{mis}\left|y_{obs}\right.)$ erreicht.

Die Annahme einer multivariaten Normalverteilung erscheint zunächst etwas restriktiv für die beobachteten Daten, die neben quantitativen Variablen auch Dummies enthalten. Dennoch hat sich diese Modellierung in vielen Studien als durchaus geeignet erwiesen, insbesondere wenn Werte nur bei den quantitativen Variablen fehlen. Wie der I-Schritt zeigt, werden die fehlenden Werte aus ihren prädiktiven bedingten Verteilungen generiert, welche im Falle der multivariaten Normalverteilung wiederum bedingte Normalverteilungen sind. Die Dummy Variablen können darin als Designmatrizen betrachtet werden, so daß diese Modellierung zulässig ist, solange die Dummies nicht selbst zu ergänzen sind. In das Schätzmodell fließen

die ausfallbehafteten Variablen Anteil der Vorleistungen, Umsatz, Anzahl der Beschäftigten und Teilzeitbeschäftigten, Anteil der Erweiterungsinvestitionen sowie Summe der Investitionen ein. Weitere verwendete Variablen sind Dummy-Variablen für Branchen, technischer Stand der Anlagen und West/Ost, diese haben keine fehlenden Werte. Die quantitativen Variablen können entweder durch Logarithmieren oder allgemeiner durch eine Box-Cox-Transformation (für Umsatz, Beschäftigtenanzahlen und Investitionssummen) oder durch eine Logit-Transformation (für die Anteile) relativ leicht hin zu einer Normalverteilung transformiert werden.

Grundsätzlich ist bei der Formulierung des Ergänzungsmodells darauf zu achten, daß der „Imputer“ nicht mehr Annahmen trifft als der Analyst. Wenn also anzunehmen ist, daß alle im Analysemodell verwendeten Variablen einen Einfluß auf den Datenausfall haben können, dann müssen diese Variablen auch im Ergänzungsmodell verwendet werden. Den Effekt der Divergenz von Analyse- und Ergänzungsmodell hat Meng (1995) untersucht und dafür den Begriff der „Uncongeniality“ kreiert.

In der vorliegenden Untersuchung wurden alle Variablen des Analysemodells auch für das Ergänzungsmodell verwendet, so daß keine Divergenzen zu erwarten sind. Nach einer Einschwingphase von 1000 Iterationen werden Ergänzungen nach 1000, 1200, 1400, 1600, 1800, 2000 Iterationen durchgeführt und gespeichert, d.h. die Anzahl der multiplen Ergänzungen wird mit m=6 gewählt. Das Programm NORM 2.03 erlaubt es auch, die Konvergenzgeschwindigkeit anhand der geschätzten Autokorrelationsfunktion der schlechtesten Anpassung graphisch zu verfolgen. Für den vorliegenden Datensatz stellt sich die Konvergenz der Markov Ketten offensichtlich schon sehr zeitig ein, so daß alle Voraussetzungen für eine erfolgreiche Imputation erfüllt sein sollten.

4.2 Das Analysemodell

Um die Auswirkung der multiplen Datenergänzung auf die Ergebnisse von Analysen zu beschreiben, werden im folgenden Schätzungen der betrieblichen Produktivität durchgeführt. Grundlage der Schätzungen soll eine Translog Produktionsfunktion sein (vgl. Berndt & Christensen 1973). Hierbei handelt es sich um eine Produktionsfunktion mit einer flexiblen funktionalen Form mit der auch Effekte 2. Ordnung bei abgeleiteten Modellen (z.B. bei der Nachfrage nach Produktionsfaktoren) analysiert werden können. Zusammen mit der verallgemeinerten Leontieff-Produktionsfunktion (vgl. Diewert 1971) wird sie daher sehr häufig bei ökonomischen Untersuchungen verwendet.

Das Translog-Modell kann als Taylor-Approximation 2. Ordnung einer allgemeinen bzw. unbekannten funktionalen Form interpretiert werden. Als Ergebnis für p Produktionsfaktoren ergibt sich (Greene 2000, 217):

$$\ln Y = \beta_0 + \sum_{j=1}^{p} \beta_j \ln X_j + \frac{1}{2} \sum_{j=1}^{p} \sum_{l=1}^{p} \gamma_{jl} \ln X_j \ln X_l + \varepsilon . \tag{10}$$

mit Y als Output, X als Produktionsfaktoren, β bzw. γ als Parameter und ε als Störgröße. Beschränkt man das Modell auf 2 Faktoren (Arbeit: L, Kapital: K) erhält man folgende Gleichung (Greene 2000, 285):

$$\ln Y = \beta_0 + \beta_1 \ln L + \beta_2 \ln K + \frac{1}{2} \gamma_1 \ln^2 L + \frac{1}{2} \gamma_2 \ln^2 K + \gamma_3 \ln L \ln K + \varepsilon . \tag{11}$$

Dieses Modell enthält die Cobb-Douglas- und die CES-Funktionen als Spezialfälle. Bei einer Cobb-Douglas Technologie gilt $\beta_4 = \beta_5 = \beta_6 = 0$, bei einer CES-Produktionsfunktion $\beta_4 = \beta_5 = \beta_6 = n \neq 0$. Zusätzlich nehmen wir für die weitere Analyse an, daß die Produktivität davon abhängt, in welcher Branche der Betrieb angesiedelt ist (BRAN) und welche Technologie dabei verwendet wird (TECH). Das Modell erweitert sich daher zu :

$$\ln Y = \beta_0 + \beta_1 \ln L + \beta_2 \ln K + \frac{1}{2} \gamma_1 \ln^2 L + \frac{1}{2} \gamma_2 \ln^2 K + \gamma_3 \ln L \ln K + \alpha \mathrm{TECH} + \lambda_g \mathrm{BRAN}_g + \varepsilon . \tag{12}$$

Desweiteren kann unterstellt werden, daß es bedeutende Produktivitätsunterschiede zwischen den alten und den neuen Bundesländern gibt. Aus diesem Grund soll die regionale Lage des Betriebes ebenfalls in die Schätzung aufgenommen werden. Durch eine einfache Dummy-Variable können nur konstante Effekte, jedoch keine Unterschiede zwischen den übrigen Parametern dargestellt werden. Da es möglicherweise differierende Elastizitäten zwischen den alten und den neuen Bundesländern gibt, wird jede erklärende Variable zusätzlich mit einer Dummy-Variable multipliziert, die eins wird, wenn der Betrieb in den alten Bundeländern liegt (WO). Dadurch kann für jede Variable überprüft werden, ob sich unterschiedliche Parameter für Ost und West ergeben, sowie wie hoch diese Differenzen sind. Das Modell verändert sich daher zu:

$$\begin{aligned} \ln Y = {} & \beta_{o0} + \beta_{o1} \ln L + \beta_{o2} \ln K + \frac{1}{2} \gamma_{o1} \ln^2 L + \frac{1}{2} \gamma_{o2} \ln^2 K + \gamma_{o3} \ln L \ln K + \alpha_o \mathrm{TECH} + \lambda_{og} \mathrm{BRAN}_g \\ & + \beta_{o0} \mathrm{WO} + \mathrm{WO} \Big\{ \beta_{w1} \ln L + \beta_{w2} \ln K + \frac{1}{2} \gamma_{w1} \ln^2 L + \frac{1}{2} \gamma_{w2} \ln^2 K + \gamma_{w3} \ln L \ln K + \alpha_w \mathrm{TECH} \\ & + \lambda_{wg} \mathrm{BRAN}_g \Big\} + \varepsilon . \end{aligned} \tag{13}$$

Aus dem IAB-Betriebspanel können einige Variablen für die empirische Überprüfung des Modells übernommen werden. Der Output Y des Betriebes soll durch die Wertschöpfung beschrieben werden. Dabei werden die Vorleistungen vom betrieblichen Umsatz abgezogen. Dies impliziert zweierlei: Zum Ersten werden nur Betriebe mit Umsatz als Maß des Geschäftsvolumens in die Analyse mit einbezogen, d.h. Banken, Versicherungen und der öffentliche Dienst können nicht berücksichtigt werden. Zweitens ist die Angabe zu den Vorleistungen mit einem sehr hohen Antwortausfall belegt. Möglicherweise entstehen daher sehr große Effekte, wenn die Daten durch das DA-Verfahren ergänzt werden. Der maximale Anteil der Vorleistungen wurde auf 99,9% beschränkt. Der Faktor Arbeit L wird durch die Anzahl der Beschäftigten abgebildet, wobei Teilzeitbeschäftigte nur mit dem Faktor 0,5 berücksichtigt werden, um das Arbeitsvolumen anzunähern. Kapital soll durch die Höhe der Investitionen approximiert werden, die nicht der Erweiterung des Anlagenbestandes dienen. Dabei steht im Hintergrund die Annahme, daß ein konstanter Teil des Kapitals „verbraucht“ wird und ersetzt werden muß, um den Kapitalbestand zu sichern. Diese Angabe ist jedoch zensiert, wenn keine Investitionen vorgenommen werden bzw. die Investitionen nicht zur Erweiterung des Kapitalbestandes dienen. Eine andere Möglichkeit, den Faktor Kapital zu approximieren, ist die Summe der Investitionen über mehrere Perioden hinweg. Der Nachteil dieses Verfahrens besteht allerdings darin, daß dabei nur Betriebe berücksichtigt werden können, die über mehrere Wellen am Panel teilgenommen haben. Einerseits besteht dann die Möglichkeit, daß ein verzerrtes Sample von „überlebenden“ Betrieben entsteht, andererseits ergibt sich eine große Reduktion an Beobachtungen. Vergleiche der Ergebnisse von Schätzungen des angegebenen Modells haben gezeigt, daß sich keine großen qualitativen Unterschiede zwischen beiden Variablen ergeben haben. Daher wird die Höhe der Investitionen, die nicht zur Erweiterung des Kapitalbestands benutzt werden, als Approximation für Kapital verwendet.

Der technische Stand (TECH) wird durch eine Dummy-Variable abgebildet, die den Wert eins annimmt, wenn der Betrieb technologisch auf dem neuesten Stand arbeitet. Zusätzlich werden bei den Schätzungen 20 Branchen berücksichtigt. Da außerdem eine Differenzierung nach alten und neuen Bundesländern vorgenommen wird, sind weitere Dummies für Bundesländer nicht im Modell enthalten. Gebietsstrukturelle Merkmale weisen bei entsprechender Spezifikation des Modells keine eigenständigen Effekte auf.

Ein Problem, das bei der Schätzung dieses Modells insbesondere durch die Verwendung der Interaktionsvariablen mit dem West/Ost-Dummy auftreten kann, ist eine Heteroskedastie aufweisende Varianz-Kovarianz-Matrix. Um dies bei den Ergebnissen zu berücksichtigen, unterstellen wir während der Schätzung eine multiplikative Form der Heteroskedastie (vgl. Greene 1998, 294f.):

(14) $V[\varepsilon_i] = e^{\alpha' z_i}$

für i=1,2,...n, mit z_i als Einflußfaktoren auf die Varianz-Kovarianz-Matrix (inkl. einer Konstanten) und α als Parametervektor. In unserem Schätzmodell nehmen wir an, daß möglicherweise alle exogenen Variablen der Regression auch einen Einfluß auf die Varianz-Struktur haben. Daher besteht z_i jeweils aus den exogenen Variablen, jedoch ohne die Interaktionsvariablen. Die Branchen sind aus Gründen der Schätzbarkeit des Modells ebenfalls nicht berücksichtigt worden.

5 Ergebnisse der betrieblichen Produktivitätsschätzungen

Beispielhaft seien die Ergebnisse der Schätzung nur unter Verwendung der verfügbaren Fälle für die West/Ost Variable in Tabelle 5 angegeben.

Tabelle 5: OLS-Schätzergebnis für die West/Ost-Variable bei Fallreduktion

Datensatz	Variable	n	Koeffizient	Std.-Fehler	t-Wert	p-Wert
Verfügbare Fälle	WESTOST	6489	0,3578	0,1150	3,1100	0,0019

Unter Annahme eines MCAR-Datenausfalls ergibt sich also ein signifikanter Produktivitätsunterschied zwischen Ost und West von über 40%[2]. Betrachtet man hingegen die Ergebnisse, die sich bei Verwendung aller Informationen aus den ergänzten Datensätzen ergeben, verringern sich die Produktivitätsunterschiede zugunsten des Ostens erheblich, siehe Tabelle 6.

Tabelle 6: Schätzergebnisse der West/Ost-Variablen auf Basis der imputierten Daten

Datensatz	Variable	n	Koeffizient	Std.-Fehler	t-Wert	p-Wert
Imputiert 1	WESTOST	10990	0,2492	0,0995	2,5055	0,0122
Imputiert 2	WESTOST	10990	0,3462	0,0975	3,5497	0,0004
Imputiert 3	WESTOST	10990	0,1968	0,0964	2,0415	0,0412
Imputiert 4	WESTOST	10990	0,3080	0,0973	3,1642	0,0016
Imputiert 5	WESTOST	10990	0,3127	0,0976	3,2032	0,0014
Imputiert 6	WESTOST	10990	0,2958	0,0986	2,9999	0,0027

Prima facie neigen möglicherweise die unproduktivere West-Betriebe eher dazu, einige Fragen nicht zu beantworten bzw. könnten es auch die produktiven Ost-Betriebe sein, deren Informationen durch die Fallreduktion verloren gehen. Da die interessierenden Schätzfunktionen theoretisch mindestens einer asymptotischen Normalverteilung folgen, können die Ml-

[2] Der marginale Effekt einer Dummyvariable kann durch $e^{\beta} - 1$ berechnet werden.

Schätzer, deren Varianzen und Konfidenzintervalle nach dem bereits beschriebenen Vorgehen berechnet werden. Die Schätzungen des Modells für den Datensatz mit den Antwortausfällen und den ergänzten Werten werden in der folgenden Tabelle dargestellt:[3]

Tabelle 7: Determinanten der betrieblichen Produktivität 2000

(abhängige Variable: betriebliche Wertschöpfung; LS-Schätzung mit multiplikativer Form der Heteroskedastie)

	Schätzung ohne Datenergänzung	Schätzung mit Datenergänzung
Konstante	10,396*** (132,018)	10,399*** (124,911)
LNN	1,015*** (26,524)	1,011*** (25,247)
LNK	0,003 (0,623)	0,003 (0,719)
LN^2N	$-0{,}03*10^{-2}$ (0,050)	0,001 (0,216)
LN^2K	0,008*** (10,436)	0,007*** (8,989)
LNN*LNK	-0,009*** (5,076)	-0,009*** (5,781)
TECH	0,102** (2,388)	0,098** (2,372)
WO	0,358*** (3,110)	0,285** (2,509)
LNK*WO	-0,005 (0,761)	-0,003 (0,592)
LNN*WO	-0,031 (0,620)	-0,016 (0,307)
LN^2N*WO	0,005 (0,670)	0,002 (0,290)
LN^2K*WO	-0,001 (1,047)	$0{,}05*10^{-2}$ (0,472)
LNN*LNK*WO	0,003 (1,256)	0,002 (0,920)
TECH*WO	-0,082 (1,418)	-0,078 (1,316)

[3] Die Schätzungen wurden mit LIMDEP V7.0 durchgeführt.

Noch Tabelle 7

Branchendummies (Referenz: Baugewerbe)		
Land- und Forstwirtschaft, Fischerei und Fischzucht	-0,382*** (4,487)	-0,338*** (3,970)
Bergbau und Gewinnung von Steinen und Erden, Energie und Wasserversorgung	0,380*** (2,962)	0,351** (2,215)
Nahrungs- und Genussmittel	-0,267*** (2,786)	-0,261*** (2,833)
Verbrauchsgüter (inkl. Holzgewerbe)	-0,370*** (5,161)	-0,338*** (4,031)
Produktionsgüter	-0,023 (0,375)	-0,006 (0,100)
Investitions- und Gebrauchsgüter	-0,051 (0,991)	-0,067 (1,296)
Handel, Instandhaltung und Reparatur	0,058 (0,976)	0,026 (0,350)
Verkehr und Nachrichtenübermittlung	-0,368*** (3,702)	-0,263** (2,132)
Kredit- und Versicherungsgewerbe	-0,305 (0,333)	0,215 (0,255)
Gastgewerbe	-0,598*** (5,514)	-0,697*** (6,820)
Erziehung und Unterricht	-0,475*** (3,720)	-0,503*** (3,456)
Gesundheits-, Veterinär- und Sozialwesen	-0,230** (2,515)	-0,293*** (3,311)
Datenverarbeitung und Datenbanken	0,231 (1,172)	0,226 (1,112)
Forschung und Entwicklung	0,310 (0,749)	0,350 (0,757)
Beratung, Markt- und Meinungsforschung, Beteiligungs-gesellschaften, Werbung	0,003 (0,023)	-0,006 (0,049)
Grundstücks- und Wohnungswesen	0,511*** (3,538)	0,553*** (3,479)
Vermietung beweglicher Sachen, sonstige DL überwiegend für Unternehmen	-0,113 (1,439)	-0,092 (1,205)
sonstige Dienstleistungen	-0,406*** (4,460)	-0,371*** (4,192)
Organisationen ohne Erwerbszweck	-1,036*** (2,633)	-0,868** (2,280)

Noch Tabelle 7

(Land- und Forstwirtschaft, Fischerei und Fischzucht)*WO	0,019 (0,127)	-0,064 (0,444)
(Bergbau und Gewinnung von Steinen und Erden, Energie und Wasserversorgung)*WO	-0,036 (0,199)	0,017 (0,086)
(Nahrungs- und Genussmittel)*WO	0,179 (1,350)	0,103 (0,773)
(Verbrauchsgüter (inkl. Holzgewerbe))*WO	0,324*** (3,224)	0,264*** (2,746)
Produktionsgüter*WO	0,100 (1,093)	0,104 (1,154)
(Investitions- und Gebrauchsgüter)*WO	0,114 (1,446)	0,133 (1,614)
(Handel, Instandhaltung und Reparatur)*WO	0,044 (0,522)	0,037 (0,370)
(Verkehr und Nachrichtenübermittlung)*WO	0,332*** (2,593)	0,199 (1,320)
(Kredit- und Versicherungsgewerbe)*WO	1,395 (1,430)	0,634 (0,726)
Gastgewerbe*WO	-0,022 (0,153)	0,059 (0,442)
(Erziehung und Unterricht)*WO	0,123 (0,554)	0,192 (0,913)
(Gesundheits-, Veterinär- und Sozialwesen)*WO	-0,049 (0,395)	0,048 (0,451)
(Datenverarbeitung und Datenbanken)*WO	0,115 (0,495)	0,120 (0,517)
(Forschung und Entwicklung)*WO	-0,333 (0,702)	-0,269 (0,526)
(Beratung, Marktforschung, Beteiligungsgesellschaften, Werbung)*WO	0,109 (0,697)	0,198 (1,351)
(Grundstücks- und Wohnungswesen)*WO	-0,273 (1,214)	-0,056 (0,270)
(Vermietung beweglicher Sachen, sonstige DL überwiegend für Unternehmen)*WO	-0,115 (1,022)	-0,038 (0,349)
(Sonstige Dienstleistungen)*WO	0,216 (1,616)	0,171 (1,270)
(Organisationen ohne Erwerbszweck)*WO	0,671 (1,415)	0,640 (1,347)

Noch Tabelle 7

Varianzfunktion:		
σ	0,948*** (25,364)	0,935*** (17,618)
LNN	-0,02 (0,547)	-0,006 (0,110)
LNK	0,006 (1,159)	-0,008 (1,261)
LN^2N	0,002 (0,321)	0,002 (0,201)
LN2K	-0,001 (1,188)	0,001 (0,548)
LNN*LNK	-0,005*** (2,708)	-0,003 (1,472)
TECH	0,195*** (4,510)	0,121 (1,458)
WO	0,129*** (3,593)	0,190*** (3,335)
Wert der LL-Funktion	-8436,82	-
χ^2 (df.) des multiplikativen Modells der Heteroskedastie	127,36*** (7)	-
Anzahl der Beobachtungen	6489	10990

Quelle: IAB-Betriebspanel 2000. Die absoluten t-Werte werden in Klammern ausgewiesen. * bzw. ** (***) signalisieren ein Signifikanzniveau von 10% bzw. 5% (1%).

Die Annahme einer homoskedastischen Verteilung der Varianz-Kovarianz-Matrix konnte bei jeder dieser Schätzungen wie auch bei der Schätzung ohne Datenergänzung zurückgewiesen werden. Die Teststatistik für einen Breusch/Pagan-Test war bei der Schätzung von einfachen OLS-Regressionen in jedem Fall signifikant. Ebenso konnte ein χ^2 – Test den Einfluß der Varianzfunktion auf das Ergebnis der Maximum-Likelihood-Schätzung in keinem Fall zurückweisen (siehe Tabelle A.5 im Anhang). Die Ergebnisse der Regressionen entsprechen den Annahmen des Modells. Die Grenzproduktivitäten der Faktoren Arbeit und Kapital hängen von der jeweiligen Nutzung der einzelnen Faktoren ab. So sinkt die Grenzproduktivität der Arbeit bei einem verstärkten Einsatz von Kapital. Dagegen hängt die Grenzproduktivität des Kapitals von der Nutzung beider Faktoren ab und steigt mit dem Kapitaleinsatz bzw. sinkt mit dem Arbeitseinsatz. Die Produktivität eines Betriebes steigt um ca. 10%, wenn er die neuesten Technologien verwendet. Ebenso wird ein konstanter Niveauunterschied zwischen den neuen und den alten Bundesländern konstatiert. Ohne die Datenergänzung liegt die Differenz bei über 40%. Die übrigen Interaktionsvariablen zeigen überwiegend insignifikante Einflüsse auf die Produktivität eines Betriebes. Lediglich einige Branchendifferenzen sind statistisch abgesichert. So ist die Produktivität im Verbrauchsgütersektor und im Bereich des Verkehrs und der Nachrichtenübermittlung im Westen um rund 40% höher als im Osten.

Auch die Varianzfunktion enthält signifikante Parameter. Neben der Konstante σ weisen auch der West/Ost-Dummy, die Technologie-Variable und der Interaktionsterm zwischen beiden Produktionsfaktoren statistisch gesicherte Werte auf. Die Ergebnisse für den ergänzten Datensatz bestätigen in weiten Teilen die Resultate der ursprünglichen Schätzung. Insbesondere bei den Branchen und der West/Ost-Variable zeigen sich jedoch bemerkenswerte Unterschiede. Für die Branchen Bergbau und Energie, Verkehr und Nachrichtenübermittlung und Organisationen ohne Erwerbszweck sinken die Produktivitätsunterschiede im Vergleich zur Referenzgruppe (Baugewerbe) um ca. 4 - 7%-Punkte. Dagegen erhöht sich der Abstand zum Baugewerbe für das Gastgewerbe, das Gesundheitswesen sowie das Grundstücks- und Wohnungswesen um mehr als 5%-Punkte. Dies deutet darauf hin, daß durch die Datenergänzung zwar andere Strukturen zum Vorschein kommen, die fehlenden Werte aber nicht zu systematischen Unter- oder Überschätzungen von Branchendifferenzen führt. Ebenso gab es keinen Vorzeichenwechsel bzw. eine Veränderung der Signifikanzen bei den Parametern. Dies gilt jedoch nicht für die Interaktionsvariablen. Der Wert des West/Ost-Dummys nimmt sehr stark ab, so daß die konstanten Produktivitätsunterschiede um mehr als 11%-Punkte abnehmen und nur noch ca. 33% betragen. Ebenso verringert sich der Unterschied zum Verbrauchsgütergewerbe in den alten Bundesländern um mehr als 8%-Punkte. Die Differenz bei der Branche Verkehr und Nachrichtenübermittlung wird sogar insig-nifikant. Es scheint so, als ob die Differenzen zwischen den alten und den neuen Bundesländern im Datensatz mit den fehlenden Werten tendenziell zu hoch ausgewiesen werden. In der Varianzfunktion haben sich ebenfalls Veränderungen ergeben. Lediglich die Konstante σ und der West/Ost-Dummy beeinflussen die heteroskedastische Struktur der Schätzfunktion. Die Auswirkungen der „Data Augmentation“ sind also bei den Ergebnissen der Regression deutlich erkennbar und führen auch zu Implikationen für die inhaltliche Interpretation der Schätzungen.

6 Zusammenfassung und Fazit

Dieses Papier beschäftigt sich mit der Möglichkeit der Ergänzung von fehlenden Werten in freiwilligen Erhebungen. Dazu wurde das Verfahren der „Data Augmentation“ vorgestellt. Dieses Verfahren beruht auf dem Prinzip der „Multiplen Imputation“ und wurde beispielhaft auf Daten des IAB-Betriebspanels von 2000 angewendet. Die Effekte der Datenergänzung wurden anhand von Schätzungen der betrieblichen Produktivität dargestellt, da hierbei eine Reihe von Variablen verwendet werden, die von Antwortausfall betroffen sind. Daher hat die Datenergänzung möglicherweise einen großen Einfluß auf die Ergebnisse der Regressionen. Der große Vorteil der multiplen Ergänzung ist daran zu sehen, daß insbesondere bei multivariaten Auswertungen zum einen keine Information durch die übliche Fallreduktion verloren geht und zum anderen sich die Sensitivität des Datenausfalls und der Ergänzung in den typischerweise weiteren MI-Konfidenzintervallen widerspiegelt. Die strenge Annahme eines

rein zufälligen Datenausfall, die bei der Fallreduktion implizit erfolgt, kann zugunsten eines weniger restriktiven bedingt zufälligen Datenausfalls „aufgeweicht“ werden. Moderne und frei verfügbare Software wie NORM 2.03 erleichtert zudem die Durchführung der Ergänzung sowie die Kombination und Berechnung der MI-Schätzergebnisse.

Die Resultate der Schätzungen mit und ohne Datenergänzung sind einander relativ ähnlich und entsprechen den Annahmen des zugrunde liegenden Modells. Data Augmentation führt also auch bei diesen recht großen Datensätzen (knapp 11000 Beobachtungen) und der hohen Anzahl der Ergänzungen (über 3500 Beobachtungen werden hinzugefügt) zu sinnvollen und interpretierbaren Ergebnissen. Dies bedeutet jedoch nicht, daß sich die Resultate beider Schätzungen exakt gleichen. Die Unterschiede zwischen den einzelnen Branchen verändern sich und besonders die Differenz zwischen den alten und den neuen Bundesländern verringert sich. Zwar gibt es auch bei den ergänzten Daten deutliche Abweichungen in beiden Landesteilen. Diese scheinen jedoch durch die fehlenden Werte überzeichnet zu werden. Dies hätte natürlich auch wirtschaftspolitische Implikationen und kann dementsprechende Maßnahmen beeinflussen. Es kann also selbst bei einem einfachen Beispiel gezeigt werden, daß ein Ergänzungsverfahren wie etwa der Data Augmentation Algorithmus die Aussagekraft von Schätzungen deutlich erhöhen und verbessern kann.

Literatur

Berndt, E. & Christensen, L. (1973): The Translog Function and the Substitution of Equipment, Structures, and Labor in U.S. Manufacturing 1929 - 68. Journal of Econometrics, 1, 81 - 114.

Brand, J.P.L. (1999) Development, Implementation and Evaluation of Multiple Imputation Strategies for the Statistical Analysis of Incomplete Data Sets, Thesis Erasmus University Rotterdam, Print Partners Ispkamp, Enschede.

Diewert, W. (1971): An Application of the Shepard Duality Problem: A Generalised Leontief Production Function. Journal of Political Economy, 79, 481 - 507.

Greene, W. (1998): LIMDEP Version 7.0 User’s Manual, Plainview NY.

Greene, W. (2000): Econometric Analysis (4th ed.), Upper Saddle River NJ.

Hartmann, J. & Kohaut, S. (2000): Analysen zu Ausfällen (Unit-Nonresponse) im IAB-Betriebspanel. Mitteilungen aus der Arbeitsmarkt- und Berufsforschung 39:4, 609 - 618.

Kölling, A. (2000): European Data Watch: The IAB-Establishment Panel. Schmollers Jahrbuch, Zeitschrift für Wirtschafts- und Sozialwissenschaften, 120:2, 291 - 300.

Li, K.H., Raghunathan, T.E. and Rubin, D.B (1991) Large-Sample Significance Levels from Multiply Imputed Data using Moment-Based Statistics and an F Reference Distribution. Journal of the American Statistical Association, 86, 416, 1065-1073.

Meng, X.L. and Rubin, D.B. (1992) Performing Likelihood ration tests with multiply-imputed data sets. Biometrica, 79, 103-111.

Meng, X.L. (1995) Multiple-Imputation Inferences with Uncongenial Source of Input (with discussion). Statistical Science, 10, 538-573.

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

Rässler (2000) Ergänzung fehlender Daten in Umfragen, Jahrbücher für Nationalökonomie und Statistik, 220/1, 64-94.

Rässler (2001) Alternative Approaches to Statistical Matching with an Application to Media Data. Als Habilitationsschrift eingereicht, Nürnberg.

Schafer, J.L. (1997) Analysis of Incomplete Multivariate Data. Chapman and Hall, London.

Schafer, J.L. (1999a) Multiple Imputation: a Primer. Statistical Methods in Medical Research, 8, 3-15.

Schafer, J.L. (1999b) Multiple Imputation under a Normal Model, Version 2. Software for Windows 95/98/NT, http://www.stat.psu.edu/~jls/misoftwa.html.

Schafer, J.L. and Olsen, M.K. (1999) Modeling and Imputation of Semicontinuous Survey Variables. Technical Report No. 00-39, The Pennsylvania State University.

Tanner, M.A. and Wong, W.H. (1987) The Calculation of Posterior Distributions by Data Augmentation (with discussion). Journal of the American Statistical Association, 82, 298, 528-550.

Van Dyk, D.A., Meng, X.L. (2001) The Art of Data Augmentation (with discussion), Journal of Computational and Graphical Statistics, 10, 1, 1-111.

Anhang

Tabelle A.1: Deskriptive Statistiken I

	Anzahl der Beobachtungen	Minimum	Maximum	Mittelwert	Standardabweichung
Ohne Datenergänzung					
LNY	6824	3,26	23,16	14,3723	2,15949
LNN	10945	-0,69	10,6	3,181	1,80981
LNK	10094	-6,91	20,21	5,1218	9,25618
LN2N	10945	0	112,37	13,3936	13,50091
LN2K	10094	0,35	408,54	111,9008	66,74636
LNNLNK	10104	-53,97	193,04	24,2189	38,70641
Datenergänzung I					
LNY	10990	3,26	22,08	14,352	2,27908
LNN	10990	-0,69	9,07	3,1757	1,80065
LNK	10990	-6,91	20,99	5,5713	9,11454
LN2N	10990	0	82,33	13,327	13,28147
LN2K	10990	0,35	440,4	114,1065	66,87838
LNNLNK	10990	-53,97	176,49	25,7313	38,58733
Datenergänzung II					
LNY	10990	3,26	21,8	14,3548	2,26883
LNN	10990	-0,69	9,07	3,1756	1,80054
LNK	10990	-6,91	21,13	5,5554	9,10852
LN2N	10990	0	82,33	13,3263	13,28147
LN2K	10990	0,35	446,42	113,82	66,64518
LNNLNK	10990	-53,97	172,74	25,6735	38,53779
Datenergänzung III					
LNY	10990	3,26	21,7	14,3655	2,26809
LNN	10990	-0,69	9,07	3,1756	1,80063
LNK	10990	-6,91	20,21	5,5737	9,10958
LN2N	10990	0	82,33	13,3265	13,28162
LN2K	10990	0,35	408,54	114,0433	66,66331
LNNLNK	10990	-53,97	172,74	25,7289	38,55053
Datenergänzung IV					
LNY	10990	3,26	21,74	14,3612	2,27094
LNN	10990	-0,69	9,07	3,1756	1,80062
LNK	10990	-6,91	20,25	5,5503	9,11737
LN2N	10990	0	82,33	13,3263	13,28065
LN2K	10990	0,35	409,87	113,925	66,80662
LNNLNK	10990	-53,97	172,74	25,6241	38,57888

Noch Tabelle A.1

Datenergänzung V					
LNY	10990	3,26	21,92	14,3508	2,27187
LNN	10990	-0,69	9,07	3,1756	1,80066
LNK	10990	-6,91	21,31	5,5671	9,1059
LN2N	10990	0	82,33	13,3263	13,28126
LN2K	10990	0,35	454,16	113,9024	66,6896
LNNLNK	10990	-55,69	182,25	25,6751	38,53987
Datenergänzung VI					
LNY	10990	3,26	21,84	14,3473	2,27901
LNN	10990	-0,69	9,07	3,1759	1,80028
LNK	10990	-6,91	21,41	5,5402	9,10947
LN2N	10990	0	82,33	13,3268	13,28115
LN2K	10990	0,35	458,23	113,6685	66,424
LNNLNK	10990	-61,47	173,44	25,5531	38,50454

Quelle: IAB-Betriebspanel 2000

Tabelle A.2: Deskriptive Statistiken II, (Anzahl, X=1)

Variablen	Ohne Ergänzung	Ergänzung I	Ergänzung II	Ergänzung III	Ergänzung IV	Ergänzung V	Ergänzung VI
WO (West=1)	6717	6717	6717	6717	6717	6717	6717
TECH (neuester Stand =1)	2418	2433	2428	2436	2430	2428	2432
Land- und Forstwirtschaft, Fischerei und Fischzucht	268	268	268	268	268	268	268
Bergbau und Gewinnung von Steinen und Erden, Energie und Wasserversorgung	155	155	155	155	155	155	155
Nahrungs- und Genussmittel	407	407	407	407	407	407	407
Verbrauchsgüter (inkl. Holzgewerbe)	773	773	773	773	773	773	773
Produktionsgüter	990	990	990	990	990	990	990
Investitions- und Gebrauchsgüter	1679	1679	1679	1679	1679	1679	1679
Baugewerbe	1431	1431	1431	1431	1431	1431	1431
Handel, Instandhaltung und Reparatur	1904	1904	1904	1904	1904	1904	1904
Verkehr und Nachrichtenübermittlung	521	521	521	521	521	521	521
Kredit- und Versicherungsgewerbe	25	25	25	25	25	25	25
Gastgewerbe	441	441	441	441	441	441	441
Erziehung und Unterricht	124	124	124	124	124	124	124
Gesundheits-, Veterinär- und Sozialwesen	616	616	616	616	616	616	616
Datenverarbeitung und Datenbanken	126	126	126	126	126	126	126
Forschung und Entwicklung	35	35	35	35	35	35	35
Beratung, Markt- und Meinungsforschung, Beteiligungsgesellschaften, Werbung	341	341	341	341	341	341	341
Grundstücks- und Wohnungswesen	136	137	137	137	137	137	137
Vermietung beweglicher Sachen, sonstige DL überwiegend für Unternehmen	581	585	585	585	585	585	585
sonstige Dienstleistungen	409	412	412	412	412	412	412
Organisationen ohne Erwerbszweck	20	20	20	20	20	20	20

Quelle: IAB-Betriebspanel 2000

Tabelle A.3: Elastizitäten der metrischen Variablen aus Tabelle 1

(nur signifikante Werte, $\partial \ln Y / \partial \ln X$)

	Schätzung ohne Datenergänzung	Schätzung mit Datenergänzung
$\varepsilon_{Y,N}$	1,015-0,009*LNK	1,011-0,009*LNK
$\varepsilon_{Y,K}$	0,008*LNK-0,009*LNN	0,007*LNK-0,009*LNN

Quelle: IAB-Betriebspanel 2000

Tabelle A.4: Marginale Effekte der Dummy-Variablen aus Tabelle 1

(nur signifikante Werte, e^{β}-1)

	Schätzung ohne Datener-gänzung	Schätzung mit Datener-gänzung
TECH	0,107	0,103
WO	0,430	0,329
Branchendummies (Referenz: Baugewerbe)		
Land- und Forstwirtschaft, Fischerei und Fischzucht	-0,317	-0,287
Bergbau und Gewinnung von Steinen und Erden, Energie und Wasserversorgung	0,462	0,421
Nahrungs- und Genussmittel	-0,234	-0,230
Verbrauchsgüter	-0,309	-0,287
Verkehr und Nachrichtenübermittlung	-0,308	-0,231
Gastgewerbe	-0,450	-0,502
Erziehung und Unterricht	-0,378	-0,395
Gesundheits-, Veterinär- und Sozialwesen	-0,205	-0,254
Grundstücks- und Wohnungswesen	0,667	0,738
sonstige Dienstleistungen	-0,333	-0,310
Organisationen ohne Erwerbszweck	-0,645	-0,580
(Verbrauchsgüter (inkl. Holzgewerbe))*WO	0,383	0,302
(Verkehr und Nachrichtenübermittlung)*WO	0,393	-

Quelle: IAB-Betriebspanel 2000

Tabelle A.5: ML-Schätzungen des Modells mit den einzelnen ergänzten Datensätzen

(abhängige Variable: betriebliche Wertschöpfung; LS-Schätzung mit multiplikativer Form der Heteroskedastie)

	I	II	III	IV	V	VI
Konstante	10,457*** (143,538)	10,412*** (147,098)	10,387*** (146,540)	10,392*** (148,593)	10,334*** (147,704)	10,412*** (144,865)
LNN	0,984*** (28,199)	0,998*** (29,071)	1,012*** (29,570)	1,038*** (30,513)	1,025*** (30,436)	1,010*** (29,665)
LNK	0,004 (0,925)	0,004 (0,866)	0,004 (0,978)	0,002 (0,536)	$0{,}01 \cdot 10^{-2}$ (0,037)	0,006 (1,577)
LN^2N	0,006 (1,101)	0,004 (0,736)	-0,001 (0,109)	$-0{,}02 \cdot 10^{-2}$ (0,033)	-0,003 (0,450)	0,001 (0,255)
LN^2K	0,007*** (10,477)	0,007*** (10,947)	0,007*** (11,405)	0,007*** (10,826)	0,008*** (12,946)	0,007*** (10,761)
LNN*LNK	-0,009*** (5,908)	-0,010*** (6,180)	-0,009*** (5,619)	-0,009*** (5,942)	-0,009*** (6,302)	-0,010*** (6,364)
TECH	0,093** (2,306)	0,123*** (3,310)	0,110*** (2,937)	0,087** (2,383)	0,086** (2,302)	0,089** (2,289)
WO	0,249** (2,505)	0,346*** (3,550)	0,197** (2,042)	0,308*** (3,164)	0,313*** (3,203)	0,296*** (2,999)
LNK*WO	-0,004 (0,759)	-0,003 (0,531)	-0,004 (0,674)	-0,005 (0,906)	$-0{,}05 \cdot 10^{-2}$ (0,085)	-0,004 (0,765)
LNN*WO	0,014 (0,312)	-0,029 (0,661)	0,013 (0,295)	-0,049 (1,138)	-0,016 (0,363)	-0,026 (0,615)
LN^2N*WO	-0,001 (0,187)	0,002 (0,339)	0,001 (0,110)	0,004 (0,551)	0,003 (0,493)	0,004 (0,517)
LN^2K*WO	$-0{,}04 \cdot 10^{-2}$ (0,539)	-0,001 (0,975)	$-0{,}04 \cdot 10^{-3}$ (0,057)	$-0{,}02 \cdot 10^{-2}$ (0,222)	-0,001 (1,596)	0,000 (0,015)
LNN*LNK*WO	0,002 (0,798)	0,003 (1,507)	0,001 (0,358)	0,003 (1,305)	0,003 (1,325)	0,002 (0,814)
TECH*WO	-0,118** (2,294)	-0,065 (1,351)	-0,055 (1,139)	-0,065 (1,366)	-0,049 (1,017)	-0,118** (2,346)
Branchendummies (Referenz: Baugewerbe)						
Land- und Forstwirtschaft, Fischerei und Fischzucht	-0,323*** (3,832)	-0,332*** (4,022)	-0,329*** (4,052)	-0,333*** (4,161)	-0,381*** (4,787)	-0,328*** (3,940)
Bergbau und Gewinnung von Steinen und Erden, Energie und Wasserversorgung	0,225* (1,797)	0,367*** (2,972)	0,346*** (2,928)	0,422*** (3,513)	0,265** (2,349)	0,484*** (4,030)
Nahrungs- und Genussmittel	-0,293*** (3,261)	-0,276*** (3,134)	-0,221** (2,536)	-0,246*** (2,871)	-0,280*** (3,277)	-0,251*** (2,805)
Verbrauchsgüter (inkl. Holzgewerbe)	-0,288*** (4,234)	-0,285*** (4,286)	-0,317*** (4,779)	-0,371*** (5,715)	-0,385*** (5,883)	-0,384*** (5,651)
Produktionsgüter	-0,013 (0,232)	0,021 (0,361)	-0,046 (0,807)	0,014 (0,254)	-0,034 (0,617)	0,020 (0,341)
Investitions- und Gebrauchsgüter	-0,092* (1,830)	-0,054 (1,097)	-0,076 (1,551)	-0,062 (1,287)	-0,060 (1,231)	-0,058 (1,151)
Handel, Instandhaltung und Reparatur	-0,032 (0,595)	0,051 (0,970)	0,051 (0,958)	0,062 (1,202)	0,060 (1,140)	-0,037 (0,687)
Verkehr und Nachrichtenübermittlung	-0,278*** (3,061)	-0,153* (1,707)	-0,281*** (3,195)	-0,202** (2,314)	-0,280*** (3,273)	-0,386*** (4,314)
Kredit- und Versicherungsgewerbe	0,341 (0,495)	0,371 (0,557)	-0,227 (0,336)	0,433 (0,648)	-0,447 (0,660)	0,819 (1,219)
Gastgewerbe	-0,668*** (7,097)	-0,655*** (7,140)	-0,747*** (8,112)	-0,751*** (8,349)	-0,683*** (7,494)	-0,681*** (7,244)
Erziehung und Unterricht	-0,634*** (5,391)	-0,498*** (4,318)	-0,381*** (3,299)	-0,525*** (4,650)	-0,470*** (4,142)	-0,510*** (4,335)

Noch Tabelle A.5

Gesundheits-, Veterinär- und Sozialwesen	-0,279*** (3,581)	-0,325*** (4,276)	-0,314*** (4,128)	-0,224*** (3,007)	-0,277*** (3,681)	-0,337*** (4,343)
Datenverarbeitung und Datenbanken	0,217 (1,019)	0,202 (1,016)	0,241 (1,201)	0,235 (1,203)	0,246 (1,232)	0,216 (1,041)
Forschung und Entwicklung	0,322 (0,670)	0,409 (0,882)	0,346 (0,761)	0,352 (0,781)	0,318 (0,718)	0,351 (0,747)
Beratung, Markt- und Meinungsforschung, Beteiligungsgesellschaften, Werbung	0,016 (0,136)	-0,011 (0,100)	-0,087 (0,777)	0,005 (0,043)	-0,024 (0,213)	0,065 (0,562)
Grundstücks- und Wohnungswesen	0,571*** (4,136)	0,605*** (4,604)	0,516*** (4,094)	0,465*** (3,627)	0,473*** (3,906)	0,688*** (5,298)
Vermietung beweglicher Sachen, sonstige DL überwiegend für Unternehmen	-0,117 (1,557)	-0,072 (0,977)	-0,069 (0,933)	-0,094 (1,310)	-0,112 (1,532)	-0,092 (1,220)
sonstige Dienstleistungen	-0,378*** (4,343)	-0,351*** (4,122)	-0,371*** (4,349)	-0,360*** (4,323)	-0,356*** (4,235)	-0,411*** (4,708)
Organisationen ohne Erwerbszweck	-0,858** (2,319)	-0,834** (2,138)	-0,903** (2,374)	-0,937** (2,511)	-0,817** (2,211)	-0,856** (2,250)
(Land- und Forstwirtschaft, Fischerei und Fischzucht)*WO	-0,057 (0,413)	-0,094 (0,691)	-0,083 (0,623)	-0,085 (0,629)	0,019 (0,141)	-0,083 (0,599)
(Bergbau und Gewinnung von Steinen und Erden, Energie und Wasserversorgung)*WO	0,074 (0,427)	0,044 (0,255)	-0,046 (0,284)	-0,067 (0,394)	0,179 (1,117)	-0,080 (0,483)
(Nahrungs- und Genussmittel)*WO	0,101 (0,865)	0,100 (0,869)	0,094 (0,836)	0,085 (0,747)	0,216* (1,912)	0,022 (0,188)
(Verbrauchsgüter (inkl. Holzgewerbe))*WO	0,292*** (3,194)	0,207** (2,289)	0,272*** (3,067)	0,260*** (2,921)	0,288*** (3,222)	0,263*** (2,873)
Produktionsgüter*WO	0,099 (1,189)	0,122 (1,464)	0,125 (1,539)	0,083 (1,002)	0,142* (1,736)	0,052 (0,620)
(Investitions- und Gebrauchsgüter)*WO	0,184** (2,519)	0,105 (1,448)	0,093 (1,311)	0,147** (2,045)	0,163** (2,259)	0,107 (1,452)
(Handel, Instandhaltung und Reparatur)*WO	0,081 (1,113)	0,001 (0,013)	0,125* (1,761)	0,014 (0,195)	-0,054 (0,747)	0,053 (0,721)
(Verkehr und Nachrichtenübermittlung)*WO	0,144 (1,283)	0,089 (0,804)	0,215** (1,979)	0,129 (1,175)	0,286*** (2,648)	0,331*** (2,978)
(Kredit- und Versicherungsgewerbe)*WO	0,527 (0,727)	0,385 (0,549)	1,013 (1,427)	0,550 (0,781)	1,330* (1,861)	$0{,}02 \cdot 10^{-3}$ (0,000)
Gastgewerbe*WO	0,001 (0,008)	-0,002 (0,019)	0,158 (1,368)	0,057 (0,496)	0,099 (0,854)	0,043 (0,361)
(Erziehung und Unterricht)*WO	0,329* (1,732)	0,205 (1,095)	0,095 (0,515)	0,254 (1,363)	0,151 (0,801)	0,120 (0,630)
(Gesundheits-, Veterinär- und Sozialwesen)*WO	0,026 (0,255)	0,054 (0,541)	0,067 (0,683)	-0,012 (0,120)	0,064 (0,642)	0,091 (0,898)
(Datenverarbeitung und Datenbanken)*WO	0,13 (0,541)	0,158 (0,698)	0,092 (0,405)	0,146 (0,653)	0,080 (0,352)	0,116 (0,491)
(Forschung und Entwicklung)*WO	-0,201 (0,388)	-0,258 (0,516)	-0,414 (0,844)	-0,379 (0,774)	-0,141 (0,292)	-0,221 (0,437)
(Beratung, Marktforschung, Beteiligungsgesellschaften, Werbung)*WO	0,194 (1,388)	0,179 (1,329)	0,273** (2,032)	0,206 (1,540)	0,219 (1,606)	0,119 (0,858)
(Grundstücks- und Wohnungswesen)*WO	-0,13 (0,684)	-0,070 (0,382)	-0,021 (0,117)	0,020 (0,111)	0,059 (0,337)	-0,198 (1,086)

Noch Tabelle A.5

(Vermietung beweglicher Sachen, sonstige DL überwiegend für Unternehmen)*WO	-0,02 (0,201)	-0,083 (0,838)	-0,033 (0,337)	-0,015 (0,154)	0,013 (0,130)	-0,089 (0,878)
(Sonstige Dienstleistungen)*WO	0,189 (1,638)	0,073 (0,643)	0,232** (2,065)	0,107 (0,951)	0,229** (2,020)	0,192* (1,656)
(Organisationen ohne Erwerbszweck)*WO	0,679 (1,461)	0,600 (1,243)	0,646 (1,380)	0,645 (1,371)	0,550 (1,186)	0,718 (1,519)
Varianzfunktion:						
σ	0,907*** (33,351)	0,885*** (33,358)	0,965*** (33,278)	0,917*** (33,428)	1,000*** (33,291)	0,935*** (33,237)
LNN	0,006 (0,226)	0,002 (0,082)	-0,019 (0,658)	-0,055* (1,945)	-0,031 (1,088)	0,062** (2,190)
LNK	-0,014*** (3,817)	-0,009** (2,365)	-0,005 (1,230)	-0,008** (2,259)	-0,001 (0,222)	-0,013*** (3,489)
LN^2N	-0,007 (1,576)	0,004 (1,023)	0,006 (1,430)	0,010** (2,429)	0,006 (1,480)	-0,009** (2,124)
LN2K	0,002*** (3,932)	0,002*** (3,465)	$-0{,}02 \cdot 10^{-2}$ (0,041)	0,001** (2,500)	-0,001* (1,737)	0,000 (0,433)
LNN*LNK	-0,002* (1,654)	-0,005*** (4,175)	-0,003** (2,482)	-0,004*** (3,077)	-0,003** (2,036)	-0,001 (0,814)
TECH	0,232*** (7,086)	0,034 (1,044)	0,100*** (3,046)	0,065** (1,993)	0,139*** (4,255)	0,156*** (4,777)
WO	0,150*** (5,342)	0,181*** (6,430)	0,139*** (4,951)	0,240*** (8,547)	0,250*** (8,909)	0,180*** (6,399)
Wert der LL-Funktion	-15674,61	-15560,79	-15322,79	-15484,35	-15497,92	-15693,53
χ^2-Test (df.) des multiplikativen Modells der Heteroskedastie	144,86*** (7)	142,72*** (7)	120,84*** (7)	147,79*** (7)	207,25*** (7)	155,80*** (7)
Anzahl der Beobachtungen	10990	10990	10990	10990	10990	10990

Quelle: IAB-Betriebspanel 2000. Die absoluten t-Werte werden in Klammern ausgewiesen. * bzw. ** (***) signalisieren ein Signifikanzniveau von 10% bzw. 5% (1%).