Multiple Imputation von fehlenden Werten mit Daten über Unterernährung und Kindersterblichkeit

Hermann Gartner *
Sandro Scheid†

Januar 2003

Zusammenfassung

In dieser Arbeit werden die Auswirkungen einer Ersetzung von fehlenden Werten auf das Ergebnis einer Regressionsanalyse untersucht. Grundlage ist eine Untersuchung von Klasen (2000) über die Unterschiede im Zusammenhang zwischen Unterernährung und Kindersterblichkeit in Afrika und Südasien. In dem Makro-Datensatz, welcher 101 Entwicklungsländer umfasst, fällt etwa ein Drittel der 273 Beobachtungen weg, da für verschiedene verwendete Variablen die Werte fehlen. Die so verloren gegangenen Informationen sollen in diese Untersuchung genutzt werden um die Schätzergebnisse zu verbessern. Hierzu wird ein Verfahren zur multiplen Imputation verwandt, in welchem mit einem Data-Augmentation-Verfahren mehrere vervollständigte Datensätze generiert werden, mit welchen dann getrennt Schätzungen durchgeführt werden. Die Ergebnisse der Schätzungen werden dann miteinander kombiniert. Durch die Auswertung mehrerer vervollständigter Datensätze wird eine höhere Effizienz der Schätzer erreicht.

Ein Vergleich von Regressionsanalysen, die mit dem vervollständigten Daten durchgeführt wurden, mit einer Complete-case-Analyse hat gezeigt, dass sich bestimmte Koeffizienten in ihrer Größenordnung geändert haben. Bei manchen Koeffizienten sind unplausible Vorzeichen aus der Complete-case Analyse verschwunden. Es ist also vorteilhaft, bei Problemen mit fehlenden Werten moderne Imputationsverfahren zu verwenden. Die wesentlichen Ergebnisse aus der Untersuchung von Klasen (2000) konnten dennoch bestätigt werden.

Durch die Ersetzung der fehlenden Werte konnten noch eine Reihe von Variablen zugänglich gemacht werden, die in den bisherigen Untersuchungen nicht verwendet wurden, da dadurch auf noch mehr Beobachtungen hätte verzichtet werden müssen.

JEL-Klassifikation: C15, O11,

*Ludwig-Maximilians-Universität München, Volkswirtschaftliches Institut
†Ludwig-Maximilians-Universität München, Statistisches Institut

1 Einführung

Unterernährung und Kindersterblichkeit sind zentrale Indikatoren, an denen der Entwicklungsstand eines Landes abgelesen werden kann (Sen, 1999). Um die knappen Ressourcen, die für Entwicklungshilfe verwendet werden, effizient einsetzen zu können, ist es erforderlich, die genauen Erklärungsfaktoren der Unterernährung und der Kindersterblichkeit zu bestimmen.

In Ländern Südasiens lässt sich eine höhere Unterernährung beobachten, als in Ländern Afrikas. Auf der anderen Seite zeigt sich, dass die Kindersterblichkeit in Afrika höher ist als in Südasien.

Es stellt sich die Frage, wie dieses Muster zu erklären ist. Sowohl Unterernährung, als auch Kindersterblichkeit werden von den ähnlichen Faktoren beeinflusst und bedingen sich gegenseitig. Zu erwarten wäre, dass in Ländern mit hoher Unterernährung auch die Kindersterblichkeit besonders hoch ist. Wenn in einem Land die Ernährungssituation schlecht ist, führt dies zu Unterernährung, aber auch zu einer höheren Anfälligkeit gegenüber Erkrankungen, was auch zu einer höheren Mortalitätsrate führt.

Grundlage des Artikels ist ein Datensatz des Seminars für empirische Wirtschaftsforschung des Volkswirtschaftlichen Instituts der LMU, der Ausschnitte der demografischen wie medizinischen Struktur in ausgewählten Entwicklungsländern beschreibt. Anhand dieses Datensatzes wurden in Klasen (2000) mögliche Ursachen für Unterernährung in Entwicklungsländern sowie Ursachen für Kindersterblichkeit in Entwicklungsländern anhand von linearen Regressionsmodellen analysiert. Der Datensatz weist in erheblichem Maße (für einzelne Variablen bis über 30% der Werte) fehlende Werte auf. Die Analysen in Klasen (2000) beschränken sich auf die Analyse der vollständig beobachteten Fälle. In die im einzelnen geschätzten Modelle gehen, je nach berücksichtigten Variablen, zwischen 105 und 199 von insgesamt 273 Beobachtungen ein. In Gartner (2000) werden die gleichen Regressionsmodelle wie in Klasen (2000) geschätzt. Im Gegensatz zu den Berechnungen in Klasen (2000) sind die fehlenden Werte durch zwei verschiedene Imputationsverfahren ersetzt. Zum einen sind sämtliche fehlenden Werte mittels einer First order regression ersetzt. Zum anderen werden die Regressionsmodelle für Datensätze berechnet, die durch Multiple Imputation des ursprünglichen Datensatzes erzeugt sind. Die Schätzungen der Regressionsmodelle anhand der vervollständigten Datensätze stimmen im wesentlichen mit den Schätzungen in Klasen (2000) überein (vgl. Gartner (2000), Seite 4-6). Die Multiple Imputation, die in Gartner (2000) verwendet wurde, geht von einer multivariaten Normalverteilung für die Variablen aus. Bisher wurde nur ein Teil der Variablen in die Imputation einbezogen. Es besteht der Wunsch, weitere Variablen mit in das Modell für die Multiple Imputation einzubeziehen. Das multiple Imputationsverfahren, das in diesem Artikel verwendet wird, soll im Gegensatz zu dem in Gartner (2000) gerechneten auch das nominale Skalenniveau einiger Variablen, die dort als normalverteilt behandelt werden, berücksichtigen. Es folgt eine Beschreibung des Datensatzes und eine Darstellung der wichtigsten Ergebnisse aus Klasen (2000). Im Weiteren ist die Multiple Imputation für das in Gartner (2000) berechnete Modell dargestellt, die anschließend für den Fall gemischter, stetiger und diskreter Variablen verallgemeinert wird. Im Anschluß ist das Modell beschrieben, mit dem hier die fehlenden Werte ersetzt werden. Es folgt dann eine Diskussion der Ergebnisse, der anhand der ersetzten Datensätze erneut durchgeführten Analysen zur Erklärung von Ursachen für Unterernährung in Entwicklungsländern sowie Ursachen für Kindersterblichkeit in Entwicklungsländern.

2 Zur Datensituation

Es liegen Daten aus insgesamt 101 Ländern vor, die aus einem Zeitraum von 1965 bis 1996 stammen. Für einen Teil der Länder liegen Beobachtungen aus mehreren Jahren vor. Der Datensatz hat also eine Panelstruktur. Welche Länder dies sind und wie viel Beobachtungen aus den jeweiligen Ländern vorliegen, ist in einer Tabelle im Anhang zu finden.

Zur Beschreibung der Unterernährungssituation dienen folgende Variablen. Der Anteil der Säuglinge mit geringem Geburtsgewicht an allen Säuglingen ist kennzeichnend für die Ernährungssituation der Mutter. Als gering gilt ein Geburtsgewicht von weniger als 2.500 Gramm.

Der Ernährungszustand der Kinder wird mit anthropometrischen Indikatoren gemessen, welche die WHO verwendet (WHO, 1983). Hierzu dient ein Z-Score. Der Z-Score ist ein standardisierter Messwert, welcher die Distanz einer Beobachtung vom Mittelwert einer Referenzgruppe misst.

Ein Maß für das Untergewicht des Individuums i im Alter t ist das gemessene Gewicht abzüglich der durchschnittlichen Gewichtes der Kinder μ_t im gleichen Alter t geteilt durch die Standardabweichung σ_t.

$$Z_i = \frac{x_i - \mu_t}{\sigma_t} \qquad (1)$$

Kinder die mehr als zwei Standardabweichungen vom Mittelwert abweichen sind leicht untergewichtig. Starkes Untergewicht liegt vor bei drei Standardabweichungen oder mehr.

Mit zwei weiteren Z-Scores lässt sich zwischen akuter und chronischer Unterernährung unterscheiden. Ein Indikator für chronische Unterernährung ist die Wachstumsverzögerung (stunting). Sie liegt vor bei einer zu geringe Größe der Kinder für ihr Alter. Kinder die dauerhaft zu wenig Nahrungszufuhr erhalten, passen ihr Wachstum an und sind somit bei gegebenem Alter kleiner.

Auszehrung (wasting) wird gemessen in Gewicht bei gegebener Körpergröße. Sie ist meist Ergebnis von kürzlich durchlebten Hunger- oder Krankheitsperioden und misst die akute Unterernährung.

Die Variablen, die hier verwendet werden (Tabelle 1), geben den Anteil der Kinder mit akuter Unterernährung (Auszehrung) bzw. mit verzögertem Wachstum an allen Kindern an.

Indikator für die Kindersterblichkeit ist die Mortalitätsrate der under 5-jährigen. Sie ist definiert als die Wahrscheinlichkeit, zwischen der Geburt und dem vollendeten 5. Lebensjahr zu sterben, bezogen auf 1000 Lebendgeburten.

Zur Erklärung der Unterernährung und Kindersterblichkeit stehen eine Reihe von Charakteristika zur Verfügung. Ein Teil der Variablen wurden in der bisherigen Untersuchung von Klasen (2000) nicht verwendet, weil dadurch noch mehr Beobachtungen verloren gegangen wären, als es ohnehin der Fall war. Einige Variablen werden für das Imputationsverfahren hinzugezogen, da sie mit den Variablen, bei welchen Werte fehlen, korreliert sind. Die Tabelle 1 gibt einen Überblick über die verwendeten Variablen.

Eine Diskussion der Ursachen der Unterernährung findet sich in UNICEF (1998). Dort wird unterschieden zwischen grundlegenden (indirekten) Ursachen und direkten Ursachen.

Zu den grundlegenden Ursachen gehören die Ressourcen, die durch die Umwelt, die technischen Möglichkeiten und durch die Menschen zur Verfügung gestellt werden. Um diese Grundlegenden Faktoren zu erfassen wird das Bruttoinlandsprodukt pro Kopf aufgenommen. Ein höheres Sozialprodukt erhöht die Ressourcen, über welche die Haushalte verfügen, und senkt so Unterernährung und Kindersterblichkeit.

Um auch zu berücksichtigen, wie das Einkommen verteilt ist, wird der Gini-Koeffizient in den Untersuchungen aufgenommen. Er misst die Ungleichverteilung des Einkommens in dem Land. In einem Land, in welchem ein großer Teil der Menschen vom Reichtum ausgeschlossen sind, ist eine höhere Unterernährung und Kindersterblichkeit zu erwarten. Auf diesen Zusammenhang hat Sen hingewiesen (z. B. Drèze und Sen (1995)).

Die Bevölkerungsdichte wirkt nachteilig auf die Unterernährung und Kindersterblichkeit weil Krankheiten sich leichter ausbreiten und weil die Landfläche relativ zur Bevölkerungsgröße knapper wird, der Preis für Nahrungsmittel und Energie ist damit höher.

Die Variable Tropenregion gibt an, welcher Flächenanteil des Landes in tropischen Gebieten liegt. In diesen Gebieten ist die Wahrscheinlichkeit einer Malariaerkrankung besonders hoch. Malaria ist in Afrika die häufigste Todesursache der unter 5-jährigen.

Direkte Einflussfaktoren sind der Zugang zu sanitären Anlagen, der Zugang zu Wasser sowie der Zugang zu Gesundheitsversorgung. Diese verbessern die Gesundheitssituation der Kinder und senken damit die Mortalitätsrate und die Unterernährung. Ein weiterer Indikator für die Gesundheitssituation ist der Anteil der Kinder, die gegen Tuberkulose bzw. gegen Tetanus, Diphtherie und Polio geimpft wurden.

Bei einer höheren Fruchtbarkeitsrate der Frauen ist zu erwarten, dass sie Unterernährung und Kindersterblichkeit erhöht, da bei einer höheren Fruchtbarkeitsrate die Zahl der Kindern, die eine Mutter zu versorgen hat, größer ist und somit die Mittel des Haushaltes sich auf mehr Kinder aufteilen. Die gesamt Fruchtbarkeitsrate, die hier verwandt wird, ist definiert als die Zahl der Kinder, welche eine Frau, die wenigstens bis zum Ende des gebärfähigen Alters lebt, im Durchschnitt gebiert.

Eine höhere Bildung der Mutter führt dazu, dass deren Kinder besser versorgt und bei Krankheit besser gepflegt werden. Es wird daher erwartet, dass sie sich positiv auf die Unterernährung und die Kindersterblichkeit auswirkt. Die Qualifikation der Mutter wird durch die Alphabetisierungsquote der Frauen erfasst.

Mehrere Variablen liegen vor, die das Stillverhalten der Mütter beschreiben. Die Stilldauer ist zum einen von Interesse, weil eine lange Stilldauer auf eine schlechte Nahrungsversorgung hindeutet. Wenn genügend andere Nahrung für das Kind vorhanden wäre, würde die Muttermilch viel früher abgesetzt.

Zum anderen führt das optimale Stillverhalten der Mütter zu einer besseren Gesundheits- und Ernährungssituation des Kindes. Dieses wird beschrieben durch den Anteil der Kinder von der Geburt bis zum 3. Monat ausschließlich gestillt wurden sowie durch den Anteil der Kinder, die vom 6. bis 9. Monat gestillt und mit Zusatznahrung gefüttert wurden.

Als weitere Variablen stehen zur Verfügung die Kalorienzufuhr pro Kopf der Bevölkerung, die Malariafälle in % der Bevölkerung sowie die durchschnittliche Lebenserwartung. Das Verhältnis der Lebenserwartung der Frauen zu jener der Männer steht für das Ausmaß der Geschlechterdiskriminierung im Land. Die Geschlechterdiskriminierung korreliert mit mehreren Faktoren, wie z.B. dem Sozialprodukt oder der Fruchtbarkeitsrate (vgl. Klasen, 1999).

Tabelle 1 gibt auch einen Überblick über die Zahl der Beobachtungen, die fehlen. Von der Variable Stilldauer fehlt über die Hälfte der Beobachtungen. Von den Unterernährungs- und Sterblichkeits-Variablen, also die zu erklären Variablen, fehlen 10-30%. Von allen Variablen, die in der Tabelle aufgelistet sind, fehlen 24 % der Werte. Dennoch fallen bei der Schätzung der Regressions-Modelle zum Teil über die Hälfte der Beobachtungen weg. Ein großer Teil der Informationen bleibt also ungenutzt.

3 Die Complete-case-Analyse in Klasen (2000)

In Klasen (2000) wird untersucht, weshalb in Südasien relativ hohe Unterernährung zu beobachten ist, die Kindersterblichkeit jedoch vergleichsweise gering ist, in Afrika südlich der Sahara aber eine höhere Kindersterblichkeit mit einer geringeren beobachteten Unterernährung einher geht.

Hierzu wird im Rahmen einer Regressionsanalyse versucht, die Erklärungsfaktoren für Variablen, welche Unterernährung und Kindersterblichkeit messen, zu finden.

Zentrale Ergebnisse der Regressionen aus Klasen (2000) sind in den Tabellen 2 und 3 zu finden. Fixe Effekte der Regionen werden durch Dummyvariablen erfasst. Unterschieden werden die Regionen: Ostasien/Pazifik, Osteurope/Zentralasien, Mittlerer Osten/Nordafrika, Karibik, Südasien, Subsahara und Lateinamerika, wobei letztere als Referenzkategorie Verwendung findet. Um zeitliche Veränderungen aufzufangen, werden die Perioden vor 1985, 1985 bis 1989, 1990 bis 1994 und 1995 und später unterschieden.

In Tabelle 2 werden Unterernährungsmodelle geschätzt. Die Ergebnisse entsprechen den theoretischen Erwartungen. In den ersten drei Spalten ist die abhängige Variable der Anteil der untergewichtigen Neugeborenen an allen Neugeborenen. Ein höheres Einkommen verringert den Anteil der Kinder mit geringem Geburtsgewicht.

In Spalte (3) wurde ein reduziertes Modell geschätzt, bei dem die Fruchtbarkeitsrate ausgelassen wurde, da diese stark von der Bildung der Frauen abhängt. Bei der Regression in Spalte (1) fallen fast 90 Beobachtungen von den potentiell 273 Beobachtungen durch fehlende Werte weg. Durch die Hinzunahme der Stilldauer in Spalte (2) fällt über die Hälfte der Beobachtungen weg. Aus Osteuropa und Zentralasien sind dann keine Beobachtungen mehr vorhanden.

Die Spalten (4) und (5) zeigen das Ergebnis eines Regressionsmodells, das die Wachstumsverzögerung, also chronische Unterernährung, erklären soll. In Regression (5) wurden die Variablen Sanitärzugang und Fruchtbarkeitsrate weg gelassen, wodurch sich die Zahl der verwendbaren Beobachtung um 22 auf 190 erhöhte. Auch hier haben die Koeffizienten das erwartete Vorzeichen. Der Regionalkoeffizinet für Südasien ist signifikant und hat den höchsten Betrag. Die gemessene Wachstumsverzögerung ist dort also besonders groß. Die letzten drei Regressionen versuchen den Anteil derer, die für ihre Größe zu wenig wiegen, zu erklären. Die Ergebnisse sind ähnlich wie bei den Regressionen zur Wachstumsverzögerung. Auch hier fallen zahlreiche Beobachtungen weg, wenn die zusätzlichen Variablen Sanitärzugang und geringes Geburtsgewicht hinzugezogen werden.

Ein Ergebnis von Klasen (2000) ist, dass die Ursachen der höheren Kindersterblichkeit in Afrika unter anderem in höheren Fruchtbarkeitsraten, geringerem Einkommen, geringerer Bevölkerungsdichte und schlechterem Impfschutz zu finden sind. Der Einfluss dieser Faktoren auf die Sterblichkeit der unter 5-jährigen wurde mit einer linearen Regression geschätzt, deren Ergebnisse in Tabelle 3 zu sehen sind.

	N	Mittelwert	Standard-abweichung	Fehlende Anzahl	Fehlende Prozent
Variablen verwendet in Klasen(2000)					
Geringes Geburtsgewicht (%)	207	13,59	8,11	66	24,18
Mod. + starkes Untergewicht (%)	188	5,76	6,04	85	31,14
Starkes Untergewicht (%)	243	20,99	15,00	30	10,99
Mod. + starke Wachstumsverzö-gerung (%)	216	30,57	15,87	57	20,88
Sterblichkeit der unter 5jährigen	220	107,64	71,12	53	19,41
Bevölkerungs-dichte	268	87,30	131,86	5	1,83
ln(BIP)	258	7,44	0,76	15	5,49
Alphabetisierung Frauen (%)	247	60,03	28,68	26	9,52
Stilldauer 20-23 Mon. (%)	133	41,17	20,34	140	51,28
Fruchtbarkeitsra-te	268	4,77	1,68	5	1,83
Sanitärzugang (%)	194	54,00	29,03	79	28,94
Impfung Tetanus (%)	205	67,99	24,21	68	24,91
zusätzliche Variablen					
Tropenregion (%)	271	0,75	0,40	2	0,73
Kalorien pro Kopf	226	2398,87	380,89	47	17,22
Stilldauer 3 Mon. (%)	153	34,86	23,47	120	43,96
Stilldauer 6-9 Mon. (%)	139	55,14	21,28	134	49,08
Malaria (%)	221	2,18	6,23	52	19,05
Gini	192	45,62	8,93	81	29,67
Lebenserwartung	230	59,99	9,72	43	15,75
Lebenserwartung Frauen/Männer	273	1,07	0,03	0	0,00
Impfung Tuberkulose (%)	201	79,51	21,46	72	26,37
Wasserzugang (%)	228	62,37	23,87	45	16,48
Gesundheitsver-sorgung (%)	154	66,91	20,65	119	43,59
Jahr	273	1988.41	6.339	0	0.00

Tabelle 1: Beschreibung des Datensatzes

Abhängige Variable	1 geringes Geburtsgewicht	2 geringes Geburtsgewicht	3 geringes Geburtsgewicht	4 leichte + starke Wachstumsverzögerung	5 leichte + starke Wachstumsverzögerung	6 starkes Untergewicht	7 leichtes + starkes Untergewicht	8 leichtes + starkes Untergewicht
Ostasien/Pazifik	2,500	2,753	2,331	6,411	6,031	2,702	15,650	15,580
	(1,03)	(1,39)	(1,03)	(2,58)	(2,59)	(0,82)	(1,73)	(1,75)
Osteurope/Zentralasien	−3,817		−4,389		−8,633	0,320		−2,905
	(2,51)		(2,46)		(3,47)	(1,55)		(2,80)
Mittlerer Osten/Nordafrika	−2,029	1,211	−2,158	0,432	−3,533	−1,409	0,164	−3,083
	(1,26)	(1,55)	(1,26)	(2,88)	(2,72)	(1,06)	(2,28)	(2,15)
Karibik	−2,023	0,874	−2,251	−8,821	−11,601	−0,360	−1,294	−1,549
	(1,75)	(2,66)	(1,74)	(3,56)	(2,95)	(1,08)	(3,99)	(2,35)
Südasien	15,304	13,305	15,001	9,683	10,620	10,103	20,642	20,677
	(1,89)	(2,04)	(1,89)	(3,99)	(3,74)	(1,44)	(4,13)	(2,96)
Subsahara	−2,866	−2,150	−2,800	−7,936	−5,079	−0,337	5,946	2,957
	(1,14)	(1,39)	(1,14)	(2,57)	(2,31)	(0,88)	(2,17)	(1,81)
d9094	0,088	−0,411	0,148	−3,737	−1,724	−0,347	0,563	−0,204
	(0,86)	(1,00)	(0,86)	(2,02)	(1,86)	(0,69)	(1,53)	(1,44)
d8589	1,616	2,150	1,837	−3,850	−2,199	−0,759	−1,136	−0,620
	(0,96)	(1,11)	(0,95)	(2,26)	(2,09)	(0,79)	(1,76)	(1,65)
pre1985	1,978	4,037	2,154	−1,686	2,765	0,520	−0,190	0,382
	(1,26)	(3,99)	(1,25)	(2,39)	(2,15)	(0,82)	(2,33)	(1,66)
Bevölkerungsdichte	0,019	0,018	0,018	0,023	0,017	0,007	0,022	0,021
	(0,00)	(0,00)	(0,00)	(0,01)	(0,01)	(0,00)	(0,01)	(0,00)
Fruchtbarkeitsrate	0,354	0,180		3,259			1,704	
	(0,38)	(0,52)		(0,84)			(0,66)	
Alphabeitisierung d. Frauen	−0,045	0,005	−0,055	0,026	−0,121	−0,066	−0,030	−0,150
	(0,02)	(0,03)	(0,02)	(0,05)	(0,04)	(0,01)	(0,04)	(0,03)
ln(BIP)	−3,136	−3,302	−3,338	−7,602	−12,191	−2,722	−2,876	−6,960
	(0,68)	(1,20)	(0,61)	(1,74)	(1,31)	(0,46)	(1,49)	(0,97)
Stillddauer20-23		0,068						
		(0,03)						
Sanitärzugang				−0,140			−0,090	
				(0,04)			(0,04)	
geringes Geburtsgewicht							0,209	
							(0,15)	
Konstante	35,429	31,378	39,231	78,426	128,598	28,639	31,172	76,564
	(6,50)	(11,64)	(4,56)	(15,60)	(9,57)	(3,43)	(13,88)	(7,19)
Angepasstes R^2	0,773	0,760	0,770	0,725	0,702	0,746	0,828	0,776
Ramsey Test P:	0,000	0,000	0,000	0,831	0,591	0,000	0,402	0,321
N	186	114	187	138	190	170	136	217

Tabelle 2: Unterernährungsmodelle mit unvollständigen Datensatz

Für alle Tabellen gilt: Ausgelassene Kategorien sind Lateinamerika und 1995+; Standardabweichungen in Klammern

Auch hier fallen durch die Hinzunahme der Regressoren geringes Geburtsgewicht und Untergewicht gegenüber dem Modell in Spalte (1) zahlreiche Beobachtungen weg. Die Hinzunahme dieser Variablen ändert nicht die Ergebnisse der Regression und deren Einfluss ist nicht signifikant. Ein Regressionsmodell, das nur die Unterernährungsvariablen enthält ist in Spalte (4) angegeben. Hier ist nur der Koeffizient Untergewicht signifikant. In Spalte (5) sind wieder die Ergebnisse eines Modells in reduzierter Form angegeben, in welchem die Fruchtbarkeitsrate ausgelassen wurde.

Eine Vermutung in Klasen (2000) zur Erklärung des beobachteten Musters der Unterernährung und der Kindersterblichkeit ist, dass die beobachtete Unterernährung durch Messfehler zu erklären ist. Er argumentiert, dass kleine Messfehler bereits zu großen Unterschieden in den gemessenen Anteilen der unterernährten Kinder führen können. Wenn keine Messfehler vorlägen, wäre zu erwarten, dass die Unterschiede in der Unterernährung durch die vorhandenen Kovariablen erklärt werden könnten. Es zeigt sich jedoch, dass es sehr große regionenspezifische Effekte gibt, die nicht durch die Regressoren erklärt werden können.

So zeigt der Koeffizient *Südasien* den höchsten Wert der Regionendummies. Diese Beobachtung ist konsistent mit der These, dass die Population in Südasien für die Körpergröße ein geringeres genetisches Potential hat, als die amerikanische Referenzgruppe, bezüglich deren die Unterernährung gemessen wird. Es ist also zu vermuten, dass die Unterernährung in Südasien nicht so groß ist, wie es die Messungen vermuten lassen.

Ein Problem bei dieser Schätzung ist, wie bereits angemerkt, dass Beobachtungen wegfallen, die vorhandenen Informationen also nicht effizient genutzt werden. Ein weiteres Problem besteht jedoch auch darin, dass die wegfallenden Beobachtungen, wie Homogenitätstests zeigen, anders verteilt sind als die verwendeten Beobachtungen. Es treten also Schichtungseffekte auf, die zu verzerrten Schätzern führen (vgl. Little und Rubin, 1986, S. 41).

4 Modell für die Multiple Imputation in Gartner (2000)

4.1 Multiple Imputation

Im Gegensatz zur Imputation von fehlenden Werten mittels First order regression (Rubin, 1976), wird bei der Multiplen Imputation die Unsicherheit der unbekannten fehlenden Werte berücksichtigt, die durch die Kovarianzen des Modells, sowie den Varianzen der Parameterschätzungen gegeben sind. Es werden anstatt eines vollständigen Datensatzes mehrere vervollständige Datensätze erzeugt. Dabei werden die fehlenden Werte aus einer im nächsten Abschnitt spezifizierten Verteilung gezogen. Dadurch erhält man anhand der verschiedenen Datensätze variierende Parameterschätzungen für ein und dasselbe Modell, die die Unsicherheit widerspiegeln. Die unterschiedlichen Ergebnisse können zu erwarteten Parameterschätzungen, sowie deren Streuungen in folgender Weise zusammengefasst werden (vgl. Schafer (1997)).

Seien $\widehat{q_t}$, $t = 1, ..., m$ die Schätzer für m vervollständigte Datensätze sowie $\widehat{U_t}$ die Varianzschätzungen der Schätzer, so ergibt sich als zusammengefasste Punktschätzung:

$$\widehat{q} = \frac{1}{m} \sum_{t=1}^{m} \widehat{q_t} \tag{2}$$

Die Varianz von $\widehat{q}$ setzt sich zusammen aus den Varianzen der Schätzer $\widehat{q_t}$ und der Varianz

	1	2	3	4	5
Abhängige Variable	Sterblichkeit der unter 5jährigen				
Ostasien/Pazifik	−15,504	−10,236	−15,656		−18,300
	(8,64)	(8,97)	(11,32)		(8,72)
Osteurope/Zentralasien	−0,619	4,399	6,471		−11,610
	(13,09)	(20,22)	(20,23)		(12,70)
Mittlerer Osten/Nordafrika	−24,417	−13,807	−14,714		−25,911
	(10,10)	(10,65)	(10,93)		(10,24)
Karibik	−23,053	−20,066	−17,403		−26,772
	(12,19)	(14,23)	(14,31)		(12,31)
Südasien	−18,682	−17,309	−28,026		−25,106
	(13,98)	(18,17)	(21,46)		(14,01)
Subsahara	4,605	19,747	20,833		7,747
	(9,19)	(9,86)	(10,01)		(9,27)
d9094	1,799	−0,922	−2,326		3,708
	(6,59)	(6,80)	(6,92)		(6,66)
d8589	12,263	1,453	2,082		15,377
	(7,46)	(7,86)	(8,12)		(7,48)
pre1985	16,007	36,488	35,071		23,424
	(10,68)	(24,30)	(24,58)		(10,52)
Bevölkerungsdichte	0,011	−0,030	−0,043		−0,001
	(0,02)	(0,03)	(0,03)		(0,02)
Fruchtbarkeitsrate	8,669	6,549	5,863		
	(3,06)	(3,24)	(3,39)		
Alphabeitisierung d. Frauen	−0,923	−0,733	−0,767		−1,154
	(0,17)	(0,18)	(0,19)		(0,15)
ln(BIP)	−34,238	−28,516	−25,713		−40,617
	(5,38)	(6,14)	(6,61)		(4,94)
Impfung		−0,300	−0,295		
		(0,16)	(0,17)		
geringes Geburtsgewicht		1,361	1,373	−0,249	
		(0,65)	(0,67)	(0,77)	
leichtes Untergewicht			0,341	2,665	
			(0,42)	(0,42)	
Konstante	374,991	334,246	314,667	48,936	475,826
	(50,76)	(61,52)	(66,88)	(8,54)	(36,53)
Angepasstes R^2	0,786	0,811	0,811	0,306	0,779
Ramsey Test P:	0,014	0,362	0,448	0,000	0,077
N	198	167	156	176	199

Tabelle 3: Sterblichkeitsmodelle mit unvollständigen Datensatz

zwischen den Punktschätzungen.

$$\widehat{Var\widehat{q}} = \frac{1}{m}\sum_{t=1}^{m}\widehat{U_t} + \sum_{t=1}^{m}(\widehat{q_t} - \widehat{q})(\widehat{q_t} - \widehat{q})^T \tag{3}$$

Die angegebenen Formeln sind für skalare und vektorwertige Schätzungen gültig. Die Ziehung der fehlenden Werte erfolgt durch den im folgenden beschriebenen Algorithmus.

4.2 Data augmentation für unvollständige multivariate normalverteilte Daten

Als Verteilung für die Beobachtungen wird im folgenden von einer multivariaten Normalverteilung ausgegangen. Dies entspricht dem im Artikel Gartner (2000) berechneten Modell. Es liegen n unabhängige Beobachtungen des normalverteilten Vektors

$$Y_i = (Y_{i1}, ..., Y_{ip}), \qquad i = 1, ..., n \tag{4}$$

vor. Der Teil der beobachteten Werte sei mit Y_{obs}, der der fehlenden Werte mit Y_{miss} bezeichnet

$$Y = (Y_{obs}, Y_{miss}). \tag{5}$$

Weiter werden die fehlenden beziehungsweise beobachteten Werte einer Beobachtung mit

$$Y_i = (Y_{i,obs}, Y_{i,miss}) \tag{6}$$

bezeichnet.

Die Verteilung von Y ist abhängig von $\Theta = (\mu, \Sigma)$, dem Mittelwertvektor und der Kovarianzmatrix von Y_i. Im Gegensatz zum bekannten EM-Algorithmus verwenden wir hier ein Bayesianisches Verfahren, bei dem für die Parameter Θ eine a-priori Verteilung festgelegt wird. Als nicht informative a-priori Verteilungen für den Erwartungswertvektor μ und die Kovarianzmatrix Σ wird üblicherweise die Verteilung mit Dichte

$$P(\Theta) \propto |\Sigma|^{-(\frac{p+1}{2})} \tag{7}$$

angenommen (vgl. Schafer (1997)). Dies entspricht einer entarteten, invertierten Wishartverteilung $W^{-1}(m, \Lambda)$ mit Parametern $m \to -1$ und $\Lambda^{-1} \to 0$ für Σ, sowie einer diffusen Verteilung für μ. Da es sich, ausgehend von einer multivariaten Normalverteilung, bei der invertierten Wishartverteilung um eine konjugierte Familie handelt, ist die a-posteriori Verteilung selbst wieder eine invertierte Wishartverteilung und kann, wie für den im folgenden beschriebenen Algorithmus notwendig, einfach gezogen werden.

Der data augmentation Algorithmus besteht nun aus zwei Schritten, die abwechselnd ausgeführt werden.

I(mputation)-Schritt:

Es wird

$$Y_{miss}^{(t+1)} \backsim P(Y_{miss}|Y_{obs}, \Theta^{(t)}) \tag{8}$$

gezogen.

P(ropability)-Schritt:

Es wird

$$\Theta^{(t+1)} \backsim P(\Theta|Y_{obs}, Y_{miss}^{(t)}) \tag{9}$$

gezogen.

$Y_{miss}^{(t)}, \Theta^{(t)}$ konvergiert gegen $P(Y_{miss}, \Theta|Y_{obs})$. Um zu beurteilen, wie oft die Schritte wiederholt werden müssen, wird in der Praxis oft die Autokorrelation der Ziehungen $Y_{miss}^{(t)}, \Theta^{(t)}$ berechnet. Es soll dann zumindest so lange gezogen werden, bis die neuen Ziehungen zu anfänglichen Ziehungen keine Korrelation mehr aufweisen. Die ersten Ziehungen ('Burn In") werden dabei nicht berücksichtigt. Dies sichert, daß für $Y_{miss}^{(t)}$ und $\Theta^{(t)}$ bereits Werte erreicht wurden, für die die Verteilung eine hohe Wahrscheinlichkeit aufweist und die Folge ($Y_{miss}^{(t)}$,$\Theta^{(t)}$) sich bereits der posteriori Verteilung genähert hat .

Als Startwerte für $\Theta^{(0)}$ kann beispielsweise die ML-Schätzung verwendet werden.

Die Ziehung von $Y_{miss}^{(t+1)}$ erfolgt schrittweise. Es werden für die einzelnen Beobachtungen Y_i die fehlenden Werte $Y_{i,miss}^{(t+1)}$ gezogen. Da Y_i normalverteilt ist, ist auch $Y_{i,miss}^{(t+1)}|Y_{i,obs}$ normalverteilt. Erwartungswert und Kovarianzmatrix können mit Hilfe des Sweep-Operators bestimmt werden. Es ergibt sich

$$E(Y_{ij}|Y_{i,obs}, \Theta^{(t)}) = a_{0j} + \sum_{k \in O(s)} a_{kj} y_{ik} \tag{10}$$

$$Cov(Y_{ij}, Y_{ik}|Y_{i,obs}, \Theta^{(t)}) = a_{jk} \tag{11}$$

für $j, k \in M(s)$, den Indizes der fehlenden Werte von Y_i. Mit a_{jk} sind die Elemente der Matrix

$$A = SWP[O(s)]\Theta^{(t)} \tag{12}$$

bezeichnet. Es wird die Matrix

$$\Theta^{(t)} := \begin{pmatrix} -1 & \mu^{(t)^T} \\ \mu^{(t)} & \Sigma^{(t)} \end{pmatrix} \tag{13}$$

an den Positionen $O(s)$ gesweept. $O(s)$ bezeichnet die Indizes der beobachteten Werte von Y_i.

Wie erwähnt wird $\Theta^{(t+1)}$ aus einer invertierten Wishartverteilung gezogen. Für die Parameter der posteriori Verteilung erhält man $m = n - 1$ und $\Lambda = (nS)^{-1}$. Dabei bezeichnet S die Stichprobenkovarianzmatrix der Y_i.

Der Artikel Gartner (2000) geht als Verteilung für die Variablen

$$y_1, ..., y_{11}, x_1, x_2, ..., x_9 \tag{14}$$

von einer multivariaten Normalverteilung aus. Dies entspricht dem eben dargestellten Modell. Bei den Variablen y_1 bis y_{11} handelt es sich um kontinuierliche Variablen. Die übrigen, vollständig beobachteten Variablen sind dichotom und kennzeichnen Regionen bzw Zeiträume. Schafer (1997) betrachtet es in einigen Fällen als legitim, kategoriale Variablen für eine multiple Imputation als normalverteilt anzunehmen. Voraussetzung ist, daß der Anteil der kategorialen Variablen an den Variablen insgesamt klein ist (vgl.

Schafer (1997), p. 147,148). Bisher ist die Abhängigkeitsstruktur, die sich durch die wiederholten Messungen in einzelnen Ländern ergibt, nicht berücksichtigt. Es besteht der Wunsch deshalb zusätzlich die Länder-Variablen mit in das Modell aufzunehmen.

5 Data augmentation für den Fall zusätzlich vollständig beobachteter kategorialer Kovariablen

Liegen stetige normalverteilte, wie kategoriale Daten vor, kann das in Schafer (1997) beschriebene general location Modell angewandt werden. In diesem Modell werden für jede der möglichen Ausprägungen der kategorialen Variablen unterschiedliche Mittelwerte für die stetigen Variablen modelliert. Berücksichtigt man in dem vorliegenden Datensatz die Länder, aus denen die Messungen stammen, als kategoriale Variablen, so wird dadurch der Abhängigkeitsstruktur, die sich durch wiederholte Messungen in einem Land ergibt, Rechnung getragen. Eine zeitliche Struktur, wie ein möglicher Trend bei einigen Variablen über die Zeit, wird durch eine zusätzliche Zeitvariable t und die quadrierte Zeit t^2 berücksichtigt. Länder, für die lediglich ein Messzeitpunkt vorliegt, werden im später beschriebenen Modell geeignet zusammengefasst. Der in Schafer (1997) beschriebene Algorithmus zur Ziehung der Datensätze vereinfacht sich bei vollständig beobachteten kategorialen Variablen. Er ist im folgenden in dieser vereinfachten Form beschrieben.

Zusätzlich zu p stetigen, multivariat normalverteilten Variablen seien q kategoriale Variablen,

$$x_i, \qquad i = 1, ..., q$$

erhoben, die Werte aus den Mengen

$$W_i \qquad i = 1, ..., q$$

annehmen. Es ergeben sich insgesamt

$$d = \prod_{i=1}^{q} |W_i|$$

mögliche Ausprägungen, die als durchnummeriert angenommen werden können. Die stetigen Variablen seien nun nach der Ausprägung der zusätzlich erhobenen kategorialen Variablen gruppiert.

$$Y_{ij}, \qquad i = 1, ..., d; j = 1, ..., n_i, \sum_i n_i = n \tag{15}$$

stellt somit die j-te Beobachtung dar, dessen kategoriale Variablen die Ausprägung i besitzen. Anders als im oben beschriebenen Modell kann der Erwartungswert der stetigen Variablen für jede Ausprägung der kategorialen Variablen verschieden sein. Von Varianzen und Kovarianzen wird angenommen, daß sie für alle Beobachtungen identisch sind. Für die Beobachtungen

$$Y_{ij} \sim N(\mu_i, \Sigma) \tag{16}$$

ergibt sich der Parametervektor

$$\Theta = (\mu_1, ..., \mu_d, \Sigma) \tag{17}$$

für den nun wieder eine priori Verteilung zu wählen ist. Im nicht informativen Fall können für die Erwartungswerte μ_i diffuse Verteilungen

$\mu_i \sim$const.

gewählt werden. Insgesamt ergibt sich dann

$$P(\Theta) \propto |\Sigma|^{-(\frac{p+1}{2})} \tag{18}$$

(vgl. Schafer (1997)). Dies entspricht der bereits oben gewählten priori Verteilung. Für die posteriori Verteilungen ergeben sich

$$\Sigma|Y \sim W^{-1}(n-d, S^{-1}) \tag{19}$$

$$\mu_i|\Sigma, Y \sim N(\widehat{\mu}_i, n_i^{-1}\Sigma) \tag{20}$$

S bezeichnet dabei die empirische Kovarianzmatrix der Residuen der Y_{ij}

Der Algorithmus selbst besteht wieder aus den Schritten:

I(mputation)-Schritt:

Ziehe $Y_{miss}^{(t+1)} \backsim P(Y_{miss}|Y_{obs}, \Theta^t)$

P(ropability)-Schritt:

Ziehe $\Theta^{t+1} \backsim P(\Theta|Y_{obs}, Y_{miss}^{(t)})$

Die Ziehung von $Y_{miss}^{(t+1)}$ erfolgt, wie oben beschrieben, für die einzelnen Beobachtungen der Reihe nach. Wendet man den Sweep Operator an, um die bedingten Verteilungen zu bestimmen, so ist $\Theta^{(t)}$ durch $\begin{pmatrix} -1 & \mu_i^{(t)^T} \\ \mu_i^{(t)} & \Sigma^{(t)} \end{pmatrix}$ zu ersetzen.

6 Ergebnisse

Eine Reihe von Variablen im hier verwendeten Datensatz können nur Werte zwischen null und eins annehmen. Die Annahme der Normalverteilung, die dem Imputationsverfahren zugrundegelegt wird, ist bei diesen Variablen nicht erfüllt. Sie wurden für das Imputationsverfahren mit der logistischen Funktion

$$f(x) = \frac{e^{-x}}{1+e^{-x}} \tag{21}$$

transformiert. Q-Q-Diagramme zeigen, dass die Normalverteilungsannahme eine vertretbare Annäherung für die transformierten Variablen darstellt. Schafer (1997) weist darauf hin, dass das Imputationsverfahren sehr robust ist und auch bei annähernder Normalverteilung noch funktioniert.

Zur Schätzung der fehlenden Werte wurde der Panel-Charakter des Datensatzes genutzt. Gibt es für ein Land zwei oder mehr Beobachtungen aus verschiedenen Jahren und fehlt beispielsweise in einem Jahr der Wert des Sozialproduktes, liegt der fehlende wahre

Wert vermutlich in der Nähe der Werte, die beobachtet wurden. Um diese Informationen zu nutzen, wurde eine kategoriale Ländervariable im Imputations-Modell aufgenommen. Die Länder, für die nur eine Beobachtung vorliegt, wurden zu Regionen, wie sie in den Unterernährungs- und Sterblichkeits-Modellen verwendet wurden, zusammengefasst.

Um die fehlenden Werte des Datensatzes zu ersetzen wurden 1000 Iterationen des oben beschriebenen Prozesses durchgeführt und dann die fehlenden Werte aus der Verteilung gezogen. Mit diesem Verfahren wurden 5 vervollständigte Datensätze generiert, mit denen verschiedene Regressionen durchgeführt wurden. Die Zusammenführung der Ergebnisse ist in den Tabellen 4 bis 7 dargestellt.

In der Spalte (1) wird jeweils noch einmal das Ergebnis der Complete-Case-Analyse präsentiert, um einen besseren Vergleich zu ermöglichen. Spalte (2) zeigt die gleiche Regression mit dem vervollständigten Datensatz. In Spalte (3) wurden weitere Variablen für die Regression hinzu genommen.

In der Tabelle 4 finden sich die Ergebnisse des Modelles für geringes Geburtsgewicht. Der Koeffizient zur Alphabetisierungsrate hat nun anders als in Klasen(2000) das erwartete Vorzeichen. Ist in einem Land die Qualifikation der Mutter höher, verringert dies den Anteil der Kinder mit zu geringem Geburtsgewicht. Der Einfluss der Fruchtbarkeitsrate erhöht sich deutlich. In Ländern, in denen die Fruchtbarkeitsrate größer ist, ist auch das Geburtsgewicht der Kinder geringer. Der Koeffizient ln(BIP) wird bei der Verwendung des vervollständigten Datensatzes kleiner. Der Einfluss des Sozialproduktes halbiert sich also nahezu. Entgegen den Erwartungen zeigen der Zugang zu Sanitäranlagen, Trinkwasser und zu Gesundheitseinrichtungen keinen signifikanten Einfluss. Auch der Ginikoeffizient ist insignifikant. Der Koeffizient Malaria ist signifikant positiv. Eine höhere Zahl von Malariafällen erhöht den Anteil der Kinder mit geringem Geburtsgewicht.

Die Modelle zum verzögertem Wachstum (Tabelle 5) kommen zu ähnlichen Ergebnissen. Auch hier wechselt das Vorzeichen bei der Alphabetisierungsrate. Der Einfluss des Zugangs zu Sanitäranlagen verringert sich im Modell mit dem vervollständigten Daten. In Tabelle 6 finden sich die Schätzung des Modells für leichtes und starkes akutes Untergewicht. Die Regression in Spalte (3) zeigt, dass der Zugang zu Gesundheitseinrichtungen einen signifikant negativen Effekt auf die Unterernährung hat.

Bei den Sterblichkeitsmodellen (Tabelle 7) ändert sich im Vergleich zur Complete-Case-Analyse das Vorzeichen des Dummies für die Region Osteuropa/Zentralasien. Allerdings ist deren Standardabweichung auch sehr hoch. Die Größenordnungen und die Vorzeichen der anderen Koeffizienten bleiben bei der Verwendung der vervollständigten Datensätze erhalten. Mit dem Multiple-Imputation-Verfahren können also die Ergebnisse von Klasen (2000) gestützt werden.

Ein ungewöhnliches Ergebnis ist, dass sich das Bestimmtheitsmaß bei den Schätzungen mit den vervollständigten Datensätzen verringert. Zu erwarten wäre eigentlich eine Erhöhung. Eine Verringerung des Bestimmtheitsmaß zeigt sich auch, wenn andere Imputationsverfahren, wie z.B. eine Regressionsanalyse, eingesetzt werden. Eine mögliche Ursache hierfür ist, dass das Datenmaterial aus Ländern, bei denen Werte fehlen, allgemein schlechter ist. Dort können die Messfehler besonders groß sein. Die Daten aus diesen Ländern sind aber in der Complete-Case-Analyse weggefallen. Mit dem vervollständigten Datensatz werden diese Beobachtungen in die Analyse mit eingeschlossen und können zu einer schlechteren Anpassung führen.

Abhängige Variable	Geringes Geburtsgewicht		
	(1)	(2)	(3)
Ostasien/Pazifik	2,753	2,766	2,643
	(1,39)	(1,27)	(1,51)
Osteurope/Zentralasien		−1,351	−4,917
		(2,14)	(5,01)
Mittlerer Osten/Nordafrika	1,211	−3,076	−2,506
	(1,55)	(1,64)	(2,49)
Karibik	0,874	−1,237	−0,768
	(2,66)	(2,45)	(2,19)
Südasien	13,305	13,808	13,490
	(2,04)	(4,95)	(4,19)
Subsahara	−2,150	−2,439	−2,681
	(1,39)	(1,38)	(1,88)
d9094	−0,411	0,169	0,710
	(1,00)	(1,57)	(1,79)
d8589	2,150	1,439	1,618
	(1,11)	(1,80)	(1,74)
pre1985	4,037	1,911	2,757
	(3,99)	(1,74)	(1,89)
Bevölkerungsdichte	0,018	0,018	0,018
	(0,00)	(0,01)	(0,00)
Fruchtbarkeitsrate	0,180	0,665	0,432
	(0,52)	(0,46)	(0,45)
Alphabetisierung d. Frauen	0,005	−0,043	−0,038
	(0,03)	(0,03)	(0,03)
ln(BIP)	−3,302	−1,862	−1,220
	(1,20)	(0,92)	(0,99)
Stillen 20-23 Mon.	0,068	0,035	0,015
	(0,03)	(0,03)	(0,04)
Sanitär			−0,036
			(0,02)
Gesundheit			−0,023
			(0,07)
Wasser			0,005
			(0,05)
Gini			−0,033
			(0,06)
Malaria			0,106
			(0,05)
Konstante	31,378	23,226	24,03
	(11,64)	(9,26)	(10,35)
N	114	273	273
R^2	0,760	0,627	0,660

Tabelle 4: Regression: Geburtsgewicht

Abhängige Variable	leichte + starke Wachstumsverzögerung		
	(1)	(2)	(3)
Ostasien/Pazifik	6,411	8,771	6,453
	(2,58)	(2,38)	(2,48)
Osteurope/Zentralasien		−0,537	−3,395
		(3,99)	(4,43)
Mittlerer Osten/Nordafrika	0,432	0,662	0,066
	(2,88)	(3,42)	(4,91)
Karibik	−8,821	−7,537	−8,663
	(3,56)	(3,34)	(3,38)
Südasien	9,683	11,965	9,859
	(3,99)	(4,15)	(4,18)
Subsahara	−7,936	−6,487	−6,885
	(2,57)	(2,32)	(2,45)
d9094	−3,737	−2,845	−2,387
	(2,02)	(1,91)	(1,93)
d8589	−3,850	−4,681	−4,034
	(2,26)	(2,25)	(2,18)
pre1985	−1,686	0,838	1,420
	(2,39)	(2,57)	(2,66)
Bevölkerungsdichte	0,023	0,025	0,027
	(0,01)	(0,01)	(0,01)
Fruchtbarkeitsrate	3,259	3,603	3,287
	(0,84)	(1,01)	(0,92)
Alphabetisierung d. Frauen	0,026	−0,012	−0,025
	(0,05)	(0,06)	(0,06)
ln(BIP)	−7,602	−9,646	−7,627
	(1,74)	(1,76)	(2,22)
Sanitär	−0,140	−0,078	−0,060
	(0,04)	(0,05)	(0,05)
Gesundheit			−0,014
			(0,05)
Wasser			−0,088
			(0,05)
Gini			−0,149
			(0,10)
Malaria			0,124
			(0,09)
Konstante	78,426	89,546	88,660
	(15,60)	(17,34)	(18,73)
N	138	273	273
R^2	0,831	0,689	0,700

Tabelle 5: Regression: Wachstumsverzögerung

Abhängige Variable	leichtes + starkes Untergewicht		
	(1)	(2)	(3)
Ostasien/Pazifik	15,650	15,465	14,794
	(1,73)	(1,90)	(1,97)
Osteurope/Zentralasien		1,375	−10,194
		(2,99)	(6,23)
Mittlerer Osten/Nordafrika	0,164	−0,218	−0,084
	(2,28)	(2,45)	(2,62)
Karibik	−1,294	−0,598	−0,289
	(3,99)	(2,77)	(2,63)
Südasien	20,642	17,172	17,791
	(4,13)	(4,46)	(3,59)
Subsahara	5,946	0,927	2,287
	(2,17)	(2,14)	(2,04)
d9094	0,563	−1,597	−0,758
	(1,53)	(1,51)	(1,66)
d8589	−1,136	−2,779	−2,146
	(1,76)	(1,70)	(1,71)
pre1985	−0,190	0,469	1,880
	(2,33)	(1,93)	(2,01)
Bevölkerungsdichte	0,022	0,021	0,021
	(0,01)	(0,01)	(0,01)
Fruchtbarkeitsrate	1,704	2,356	1,879
	(0,66)	(0,74)	(0,66)
Alphabetisierung d. Frauen	−0,030	−0,044	−0,035
	(0,04)	(0,04)	(0,04)
ln(BIP)	−2,876	−4,838	−3,647
	(1,49)	(1,50)	(1,74)
geringes Geburtsgewicht	−0,090	−0,052	0,137
	(0,04)	(0,03)	(0,11)
Sanitär	0,209	0,242	−0,057
	(0,15)	(0,16)	(0,04)
Gesundheit			−0,114
			(0,03)
Wasser			0,017
			(0,04)
Gini			−0,207
			(0,07)
Malaria			0,082
			(0,09)
Konstante	31,172	44,862	54,118
	(13,88)	(13,12)	(14,04)
N	136	273	273
R^2	0,830	0,741	0,771

Tabelle 6: Regression: Untergewicht

Abhängige Variable	Mortalitätsrate bis 5j.		
	(1)	(2)	(3)
Ostasien/Pazifik	−15,504	−10,586	−20,549
	(8,64)	(8,93)	(10,86)
Osteurope/Zentralasien	−0,619	10,600	−11,505
	(13,09)	(13,69)	(22,56)
Mittlerer Osten/Nordafrika	−24,417	−12,592	−8,668
	(10,10)	(11,02)	(20,68)
Karibik	−23,053	−17,208	−18,651
	(12,19)	(12,64)	(13,64)
Südasien	−18,682	−10,730	−21,173
	(13,98)	(15,17)	(16,46)
Subsahara	4,605	3,869	3,359
	(9,19)	(11,00)	(11,85)
d9094	1,799	4,080	8,685
	(6,59)	(7,34)	(7,55)
d8589	12,263	11,537	15,138
	(7,46)	(8,35)	(8,13)
pre1985	16,007	21,965	28,536
	(10,68)	(10,64)	(11,71)
Bevölkerungsdichte	0,011	0,003	0,013
	(0,02)	(0,02)	(0,03)
Fruchtbarkeitsrate	8,669	10,262	8,189
	(3,06)	(3,56)	(3,33)
Alphabetisierung d. Frauen	−0,923	−0,968	−0,942
	(0,17)	(0,18)	(0,19)
ln(BIP)	−34,238	−37,220	−26,846
	(5,38)	(5,36)	(6,31)
Sanitär			−0,111
			(0,12)
Gesundheit			−0,141
			(0,39)
Wasser			−0,343
			(0,20)
Gini			−0,612
			(0,40)
Malaria			0,601
			(0,39)
Konstante	374,991	388,590	380,145
	(50,76)	(54,70)	(52,96)
N	198	273	273
R^2	0,786	0,764	0,748

Tabelle 7: Regression: Kindersterblichkeit

7 Schlussfolgerung

In dieser Arbeit wurde ein Makro-Datensatz, in dem eine Reihe von Werten fehlten, mit einem Multiple-Imputation-Verfahren vervollständigt. Ein Vergleich von Regressionsanalysen, die mit dem vervollständigten Daten durchgeführt wurden, mit einer Complete-case-Analyse hat gezeigt, dass sich bestimmte Koeffizienten in ihrer Größenordnung geändert haben. Bei manchen Koeffizienten sind unpdlausible Vorzeichen aus der Complete-case Analyse verschwunden. Es ist also vorteilhaft, bei Problemen mit fehlenden Werten moderne Imputationsverfahren zu verwenden, welche die in den Daten vorhanden Informationen besser ausnutzen können. Schichtungseffekte, die gegebenenfalls bei einer Complete-case-Analyse auftreten und zu verzerrten Schätzern führen, können vermieden werden. Die wesentlichen Ergebnisse aus der Untersuchung von Klasen (2000) konnten dennoch bestätigt werden.

Durch die Ersetzung der fehlenden Werte konnten noch eine Reihe von Variablen zugänglich gemacht werden, die in den bisherigen Untersuchungen nicht verwendet wurden, da dadurch auf noch mehr Beobachtungen hätte verzichtet werden müssen. So wurden noch Informationen über die Einkommensverteilung, den Zugang zu Wasser und Gesundheitseinrichtungen verfügbar. Ein Ergebnis ist, dass eine bessere Gesundheitsversorgung die akute Unterernährung eines Landes verringert.

Literatur

Drèze, J. und **Sen, A.** (1995). *India – Economic Development and Social Opportunity.* Oxford University Press, Oxford.

Gartner, H. (2000). Die Ersetzung fehlender Werte: Ein Test alternativer Methoden mit Makrodaten. SFB 386, Ludwig-Maximilians-Universität München, Discussion Paper 216.

Klasen, S. (1999). Does gender inequality reduce growth and development? Evidence from cross-country regressions. World Bank Policy Research Report - Working Paper No. 7.

Klasen, S. (2000). Malnourished and surviving in South Asia, better nourished and dying in Africa: What can explain this puzzle? SFB 386, Ludwig-Maximilians-Universität München, Discussion Paper 214.

Little, R. J. A. und **Rubin, D. B.** (1986). *Statistical Analysis with Missing Data.* John Wiley & Sons, New York.

Rubin, D. B. (1976). Inference and missing data. *Biometrica*, 63(3), 581–592.

Schafer, J. L. (1997). *Analysis of Incomplete Multivariate Data.* Chapman & Hall, Londod.

Sen, A. (1999). *Development as Freedom.* Oxford University Press, Oxford.

UNICEF (1998). *The State of the World's Children 1998: Focus on Nutrition.* Oxford University Press, Oxford.

WHO (1983). *Measuring Change in Nutritional Status.* WHO, Geneva.

Land	*Beob.*	Land	*Beob.*	Land	*Beob.*
Algeria	2	Gabon	1	Niger	2
Argentina	1	Gambia	4	Nigeria	2
Azerbaijan	1	Ghana	3	Papua New Guinea	2
Bahrein	1	Guatemala	2	Pakistan	4
Bangladesh	5	Guinea	1	Panama	3
Barbados	1	Guyana	3	Paraguay	1
Benin	1	Haiti	3	Peru	4
Bolivia	9	Honduras	3	Philippines	6
Botswana	3	India	4	Romania	1
Brasil	3	Indonesia	2	Russia	2
Burkina	1	Iran	1	Rwanda	4
Burundi	1	Iraq	1	Senegal	4
Central Afr. Rep.	2	Jamaica	5	Seychelles	1
Cameroon	1	Jordan	2	Sierra Leone	3
CapeVerde	2	Kazakstan	1	South Africa	1
Chile	17	Kenya	3	Sri Lanka	3
China	2	Kyrghistan	1	Sudan	8
Colombia	5	Laos	2	Swaziland	1
Comoros	2	Lebanon	1	Syria	1
Congo	1	Lesotho	3	Tanzania	2
Costa Rica	7	Liberia	2	Thailand	1
Croatia	3	Madagascar	4	Togo	4
Czech Rep.	1	Malawi	4	Trinidad	2
Côte d'Ivoir	2	Malaysia	5	Tunisia	3
Dem. Rep. Congo	1	Mali	2	Turkey	2
Djibouti	1	Mauritania	3	Uganda	2
Dominican Rep.	2	Mauritius	2	Uruguay	3
Ecuador	1	Mexico	4	Uzbekistan	1
Egypt	6	Mongolia	1	Venezuela	4
El Salvador	3	Mozambique	2	Vietnam	3
Equatorial Guinea	1	Myanmar	6	Yemen	7
Eritrea	2	Namibia	1	Zambia	3
Ethiopia	2	Nepal	2	Zimbabwe	2
Fiji	1	Nicaragua	3		
Total					273

Tabelle 8: Liste der im Datensatz vorhandenen Länder