Eine allgemeine Formel zur Anpassung an Randtabellen

Gabler, Siegfried

Veröffentlichungsversion / Published Version
Zeitschriftenartikel / journal article

Eine allgemeine Formel zur Anpassung an Randtabellen

Von Siegfried Gabler

Die Frage der Anpassung von Zellhäufigkeiten an bekannte Randhäufigkeiten nach der Durchführung einer Umfrage beschäftigt die Sozialwissenschaftler schon lange. Eine allgemein übliche Vorgehensweise dieser Reparaturtechnik liefert der Iterative Proportional Fitting, kurz IPF-Algorithmus. Der Nachteil dieses Verfahrens besteht darin, daß die Anpassung unabhängig vom interessierenden Merkmal vollzogen wird. Nur die Anpassungsmerkmale spielen eine Rolle. Läßt sich der (unbekannte) Häufigkeitsvektor des interessierenden Merkmals als Element eines gegebenen Parameterraumes lokalisieren, besteht die Möglichkeit, diese Kenntnis in die Gewichtung einzubeziehen. An Hand der Anpassung des ALLBUS 88 an den Mikrozensus 87 wird die Vorgehensweise beispielhaft verdeutlicht.

1. Einleitung

Der ALLBUS liefert bekanntlich für bestimmte Merkmale vom Mikrozensus abweichende Häufigkeiten. Einen Vergleich beider Bevölkerungsumfragen hat Hartmann (1990) gegeben. So enthält der ALLBUS beispielsweise weniger Personen in hohem Alter und weniger Einpersonenhaushalte als der amtliche Mikrozensus. Schafft eine Anpassung der ALLBUS Häufigkeiten durch Gewichtung an die Mikrozensus Häufigkeiten Abhilfe? In denselben ZUMA-Nachrichten weist Rothe (1990) anhand konkreter Fälle auf die Gefahren der Verwendung von Globalgewichten hin. So erzielt eine Globalgewichtung der Zellen des ALLBUS 86 mittels IPF Anpassung an die Merkmale Alter, Geschlecht, Bildung und Stellung im Beruf aus dem Mikrozensus 85 gegenüber der Nicht-Gewichtung zwar eine Verbesserung bei den Merkmalen Branche, Wochenarbeitszeit, Kinderzahl unter fünfzehn Jahren, jedoch eine Verschlechterung bei den Merkmalen Familienstand, Arbeitslosigkeit und Haushaltseinkommen. Für jedes einzelne Merkmal müßte in die Gewichtung der Ausfallmechanismus einfließen, von dem man in der Regel jedoch keine oder nur vage Kenntnisse besitzt. Im folgenden soll ein Modell vorgestellt werden, das die Möglichkeit bietet, Zusatzinformationen über das interessierende Merkmal bei der Konstruktion von Gewichten zu berücksichtigen. Der IPF- Algorithmus besitzt diese Flexibilität nicht.

2. Das Modell

Die übliche Vorgehensweise bei den Gewichtungsprozeduren wäre, die Häufigkeiten des ALLBUS so zu gewichten, daß die gewichteten Häufigkeiten dieselben Randverteilungen liefern wie die Randverteilungen des Mikrozensus. Das bekannteste und zugleich älteste Verfahren ist der Iterative Proportional Fitting Algorithmus von Deming/Stephan

(1940), der auch als Raking bezeichnet wird. Andere Methoden sind die Maximum Likelihood Methode bei uneingeschränkter Zufallsauswahl, die Minimum χ^2 Methode und die Kleinst Quadrate Methode. Einen Vergleich dieser Methoden und die zugrundeliegenden Modellvorstellungen haben Causey (1983) und Little/Wu (1991) gegeben. Die resultierenden Gewichtungen sind Globalgewichte, also unabhängig von den eigentlichen Untersuchungsmerkmalen konstruiert. In aller Regel interessieren nicht so sehr die einzelnen Zellhäufigkeiten, sondern Randverteilungen bestimmter Merkmale. So soll etwa die Verteilung des Haushaltseinkommens oder des Familienstandes usw. geschätzt werden. Für die Anpassungsvariablen stimmen die Verteilungen im ALLBUS und Mikrozensus nach der Gewichtung exakt überein, jedoch nur mehr oder weniger gut die Verteilungen anderer Merkmale. Ziel dieser Arbeit ist es, in die Gewichtung neben den Anpassungsvariablen auch Kenntnisse über das interessierende Merkmal einzubeziehen. Dies soll bewirken, daß sich auch die Verteilung des interessierenden Merkmals im ALLBUS „möglichst wenig" von seiner Verteilung im Mikrozensus unterscheidet. Dabei ist die Bedeutung von „möglichst wenig" zu präzisieren. Formelmäßig wird dies im mathematischen Anhang getan. Verbal läßt sich die Idee beispielhaft wie folgt beschreiben: Wir wissen von den interessierenden Merkmalsausprägungen nur, daß ihre „Varianz" einen bestimmten Wert nicht überschreitet. Die Gewichtung wird nach der Minimax-Regel nun so konstruiert, daß bei dieser Kenntnis der dann noch maximal mögliche quadrierte Abstand zwischen gewichteten ALLBUS Häufigkeiten des interessierenden Merkmals und den Mikrozensuswerten möglichst klein wird, und die exakte Anpassung an vorgegebene Merkmale erhalten bleibt. Erstaunlicherweise läßt sich zeigen, daß jede Gewichtung ein Spezialfall dieses Modells ist. Insbesondere ist die IPF-Lösung in diesem Modell enthalten. Die verschiedenen Spezifikationen des Varianzbegriffes ermöglichen verschiedene Gewichtungen. Die zur Spezifikation des Modells benötigten Werte werden durch eine Diagonalmatrix festgelegt (siehe Anhang).

3. Minimax-Gewichtung beim ALLBUS 88

Als Beispiel für eine Minimax-Gewichtung mit realen Daten dienen uns Variablen des ALLBUS 88. Als vorgegebene Randverteilungen verwenden wir den Mikrozensus 87. Es soll hier nicht weiter auf Probleme der Vergleichbarkeit beider Umfragen eingegangen werden. Der interessierte Leser sei auf den Artikel von Hartmann (1990) verwiesen. Uns geht es hier in erster Linie um einen Vergleich mit dem IPF-Algorithmus. Die Berechnungen wurden auf einem 486er PC mittels in GAUSS2.1 geschriebenen Programmen durchgeführt. Die Berechnung der Ergebnisse benötigte für die Minimax-Gewichtung weniger CPU-Zeit als für die IPF-Gewichtung. Die zur Spezifikation benötigten Werte könnten aus Vergangenheitsgewichten ALLBUS 86 zu Mikrozensus 85 bestimmt werden, was nicht unbedingt vernünftig sein muß, wie in anderem Zusammenhang Arminger (1990) zeigt. Die Güte dieser Vorgehensweise entsprach etwa der des IPF-Algorithmus. Bei der Hochrechnung der Zellhäufigkeiten von den Anpassungsvariablen auf interessierende- plus Anpassungsvariablen ließen sich ebenfalls Vergangenheitsverhältnisse einbauen. Im weiteren wurde dies jedoch nicht verwendet. Vielmehr werden zu den wahren (normalerweise un-

bekannten) Gewichten die Projektionen auf den von den Spalten der Anpassungsvariablen erzeugten orthogonalen Unterraum berechnet und daraus die Spezifikation des Modells bestimmt. Die betrachteten Variablen waren das Geschlecht, das Alter, der höchste Bildungsabschluß und die Stellung im Beruf. Je zwei davon sind Anpassungsvariablen, für die beiden anderen Merkmale werden jeweils die Randverteilungen berechnet. Als Güte der Anpassung wird Pearsons χ^2 angegeben. Im einzelnen ergaben sich folgende Tabellen:

Anpassungsvariablen : Stellung im Beruf Bildung

Geschlecht	ALLBUS 88	ALLBUS 88 gewichtet		Mikrozensus 87
		Minimax	IPF	
männlich	0.444299	0.439649	0.441299	0.465136
weiblich	0.555701	0.560351	0.558701	0.534864
	0.001745	0.002611	0.002284	Pearsons χ^2

Anpassungsvariablen : Alter Stellung im Beruf

Geschlecht	ALLBUS 88	ALLBUS 88 gewichtet		Mikrozensus 87
		Minimax	IPF	
männlich	0.444299	0.479366	0.484241	0.465136
weiblich	0.555701	0.520634	0.515759	0.534864
	0.001745	0.000814	0.001467	Pearsons χ^2

Anpassungsvariablen : Alter Bildung

Geschlecht	ALLBUS 88	ALLBUS 88 gewichtet		Mikrozensus 87
		Minimax	IPF	
männlich	0.444299	0.408713	0.407207	0.465136
weiblich	0.555701	0.591287	0.592793	0.534864
	0.001745	0.012796	0.013489	Pearsons χ^2

Anpassungsvariablen : Geschlecht Alter

Stellung im Beruf	ALLBUS 88	ALLBUS 88 gewichtet		Mikrozensus 87
		Minimax	IPF	
Selbständig	0.052425	0.054957	0.053705	0.060203
Beamte	0.047510	0.049350	0.048158	0.051757
Angestellte	0.216252	0.208061	0.208379	0.218957
Arbeiter	0.127785	0.132924	0.129198	0.184370
nicht erwerbstätig	0.556029	0.554708	0.560561	0.484713
	0.029246	0.025574	0.029842	Pearsons χ^2

Anpassungsvariablen : Geschlecht Bildung

Stellung im Beruf	ALLBUS 88	ALLBUS 88 gewichtet		Mikrozensus 87
		Minimax	IPF	
Selbständig	0.052425	0.048638	0.049389	0.060203
Beamte	0.047510	0.039693	0.040599	0.051757
Angestellte	0.216252	0.182498	0.182994	0.218957
Arbeiter	0.127785	0.150530	0.147577	0.184370
nicht erwerbstätig	0.556029	0.578641	0.579441	0.484713
	0.029246	0.035517	0.036110	Pearsons χ^2

Anpassungsvariablen : Alter Bildung

Stellung im Beruf	ALLBUS 88	ALLBUS 88 gewichtet		Mikrozensus 87
		Minimax	IPF	
Selbständig	0.052425	0.048737	0.048538	0.060203
Beamte	0.047510	0.039998	0.039447	0.051757
Angestellte	0.216252	0.189658	0.189028	0.218957
Arbeiter	0.127785	0.151097	0.140293	0.184370
nicht erwerbstätig	0.556029	0.570510	0.582694	0.484713
	0.029246	0.029967	0.039622	Pearsons χ^2

Anpassungsvariablen : Geschlecht Alter

Bildung	ALLBUS 88	ALLBUS 88 gewichtet		Mikrozensus 87
		Minimax	IPF	
Hauptschule ohne Lehre	0.176933	0.185610	0.182727	0.299905
Hauptschule mit Lehre	0.394823	0.400591	0.403736	0.374674
Realschule	0.230668	0.222085	0.223102	0.191260
Abitur	0.112385	0.105723	0.105360	0.068576
Hochschulabschluß	0.085190	0.085991	0.085075	0.065585
	0.093475	0.076790	0.078862	Pearsons χ^2

Anpassungsvariablen : Geschlecht Stellung im Beruf

Bildung	ALLBUS 88	ALLBUS 88 gewichtet		Mikrozensus 87
		Minimax	IPF	
Hauptschule ohne Lehre	0.176933	0.173881	0.170623	0.299905
Hauptschule mit Lehre	0.394823	0.405698	0.407977	0.374674
Realschule	0.230668	0.229216	0.230109	0.191260
Abitur	0.112385	0.105097	0.105485	0.068576
Hochschulabschluß	0.085190	0.086107	0.085806	0.065585
	0.093475	0.088930	0.092683	Pearsons χ^2

Anpassungsvariablen : Alter Stellung im Beruf

Bildung	ALLBUS 88	ALLBUS 88 gewichtet		Mikrozensus 87
		Minimax	IPF	
Hauptschule ohne Lehre	0.176933	0.184063	0.182231	0.299905
Hauptschule mit Lehre	0.394823	0.412893	0.414783	0.374674
Realschule	0.230668	0.228821	0.222806	0.191260
Abitur	0.112385	0.093214	0.094972	0.068576
Hochschulabschluß	0.085190	0.081008	0.085208	0.065585
	0.093475	0.068500	0.071701	Pearsons χ^2

Anpassungsvariablen : Geschlecht Stellung im Beruf

Alter	ALLBUS 88	ALLBUS 88 gewichtet		Mikrozensus 87
		Minimax	IPF	
18 und 19 Jahre	0.035387	0.035411	0.035021	0.038355
20-24	0.107798	0.111677	0.111661	0.104341
25-29	0.117955	0.121542	0.122189	0.093823
30-34	0.094037	0.098353	0.099033	0.081392
35-39	0.086501	0.092266	0.092743	0.077276
40-44	0.069790	0.073714	0.073292	0.069176
45-49	0.080603	0.086604	0.085606	0.097467
50-54	0.074050	0.076146	0.076359	0.083028
55-59	0.069790	0.070211	0.070547	0.077216
60-64	0.073394	0.067178	0.067062	0.073182
65-69	0.069790	0.061259	0.061070	0.058571
70 Jahre und älter	0.120904	0.105639	0.105418	0.146171
	0.020742	0.029943	0.031079	Pearsons χ^2

Anpassungsvariablen : Geschlecht Bildung

Alter	ALLBUS 88	ALLBUS 88 gewichtet		Mikrozensus 87
		Minimax	IPF	
18 und 19 Jahre	0.035387	0.036914	0.036021	0.038355
20-24	0.107798	0.095342	0.094051	0.104341
25-29	0.117955	0.101972	0.101069	0.093823
30-34	0.094037	0.085052	0.084047	0.081392
35-39	0.086501	0.077507	0.076999	0.077276
40-44	0.069790	0.066055	0.066233	0.069176
45-49	0.080603	0.079444	0.079223	0.097467
50-54	0.074050	0.078443	0.079808	0.083028
55-59	0.069790	0.077797	0.078342	0.077216
60-64	0.073394	0.079537	0.079607	0.073182
65-69	0.069790	0.078952	0.079744	0.058571
70 Jahre und älter	0.120904	0.142985	0.144854	0.146171
	0.020742	0.013147	0.013715	Pearsons χ^2

Anpassungsvariablen : Stellung im Beruf Bildung

Alter	ALLBUS 88	ALLBUS 88 gewichtet		Mikrozensus 87
		Minimax	IPF	
18 und 19 Jahre	0.035387	0.034746	0.033900	0.038355
20-24	0.107798	0.100033	0.100016	0.104341
25-29	0.117955	0.106671	0.107789	0.093823
30-34	0.094037	0.091936	0.090778	0.081392
35-39	0.086501	0.086488	0.085938	0.077276
40-44	0.069790	0.069286	0.071123	0.069176
45-49	0.080603	0.089959	0.088396	0.097467
50-54	0.074050	0.085892	0.085113	0.083028
55-59	0.069790	0.078552	0.079355	0.077216
60-64	0.073394	0.068337	0.068953	0.073182
65-69	0.069790	0.066360	0.066611	0.058571
70 Jahre und älter	0.120904	0.121740	0.122028	0.146171
	0.020742	0.010881	0.011175	Pearsons χ^2

Wie aus den letzten Zeilen der Tabellen abzulesen ist, ist die Güte der Anpassung bis auf die erste Tabelle bei der Minimax-Gewichtung besser als bei der IPF-Gewichtung. Es ist jedoch nicht ganz auszuschließen, daß in einzelnen Fällen die ungewichteten ALLBUS Werte besser als die gewichteten ALLBUS Werte sind.

4. Schlußbetrachtung und Zusammenfassung

Sollen in die Gewichtung neben den Anpassungsvariablen auch Kenntnisse über das Untersuchungsmerkmal einfließen, bietet die Minimax-Gewichtung dazu die Möglichkeit. Die IPF-Gewichtung ist nur ein Spezialfall einer Minimax-Gewichtung und daher schlecht, wenn die Modellvorstellung nicht den Gegebenheiten entspricht. Gute Minimax-Gewichtungen hängen von einer guten Spezifikation der Streuung der Ausprägungen des Untersuchungsmerkmals ab. Da die Spezifikation für jede Untersuchungsvariable getrennt vorgenommen werden kann, liefert die Minimax-Gewichtung ein flexibles Handwerkszeug der Gewichtungspraxis als Reparaturwerkzeug. Gewichten oder nicht Gewichten, das ist nicht mehr die Frage, sondern wie gut oder wie schlecht ist meine Spezifikation?

Anmerkung
Für die Datenbeschaffung und -aufbereitung danke ich Herrn Bernhard Schimpl-Neimanns aus der Mikrodatenabteilung.

Literatur

Arminger, G., 1990: Pflicht- versus Freiwilligenerhebung im Mikrozensus. Allgemeines Statistisches Archiv 74:161-187.

Causey, D., 1983: Estimation of Proportions for Multinomial Contingency Tables Subject to Marginal Constraints. Communications in Statistics - Theory and Methods 12:2581-2587.

Deming, E./Stephan, F., 1940: On a Least Squares Adjustment of a Sampled Frequency Table when the Expected Marginal Totals are known. The Annals of Mathematical Statistics 11:427-444.

Gabler, S., 1990a: An Identity for Reflexive g-inverses in the Context with BLU Estimators. Statistische Hefte 31:225-231.

Gabler, S., 1990b: Minimax Solutions in Sampling from Finite Populations. New York: Springer Verlag.

Hartmann, P., 1990: Wie repräsentativ sind Bevölkerungsumfragen? Ein Vergleich des ALLBUS und des Mikrozensus. ZUMA-Nachrichten 26:7-30.

Little,J.A./Wu, M., 1991: Models for Contingency Tables with Known Margins when Target and Sampled Populations Differ. Journal of the American Statistical Association 86:87-95.

Merz, J., 1983: Die konsistente Hochrechnung von Mikrodaten nach dem Prinzip des minimalen Informationsverlustes. Allgemeines Statistisches Archiv 67:342-366.

Rao, C.R./ S.K.Mitra, 1971: Generalized Inverse of Matrices and its Applications. New York: John Wiley & Sons, Inc.

Rothe, G., 1990: Wie (un)wichtig sind Gewichtungen? Eine Untersuchung am ALLBUS 1986. ZUMA-Nachrichten 26:31-55.

5. Mathematischer Anhang

Es sei $\mathcal{C}(N,K)$ die Menge aller Matrizen mit N Zeilen und K Spalten. Das Tensor- oder Kroneckerprodukt $A.\star.B$ ist definiert durch $A.\star.B = (a_{ij}B)_{i,j}$.
Für $A,B \in \mathcal{C}(N,K)$, $t \in \mathcal{C}(N,1)$ ist die elementweise Multiplikation bzw. Division definiert durch

$$A.\star B = (a_{ij}b_{ij})_{i,j} \qquad A.\star t = (a_{ij}t_i)_{i,j}$$

$$A./B = (\frac{a_{ij}}{b_{ij}})_{i,j} \qquad A./t = (\frac{a_{ij}}{t_i})_{i,j}.$$

Dann gilt

Lemma 1 .

a) $(AB).\star t = (t.\star A)B$ für $A \in \mathcal{C}(N,M), B \in \mathcal{C}(M,H), t \in \mathcal{C}(N,1)$

b) $A.\star b = BA, \quad A./b = B^{-1}A$ für $A \in \mathcal{C}(N,M), b \in \mathcal{C}(N,1)$ und $B = diag(b)$.

$diag(b)$ ist die Diagonalmatrix, deren Diagonalelemente aus den Komponenten des Vektors b bestehen.

Lemma 2 .

Es sei m ein Spaltenvektor und $M = diag(m)$ invertierbar. $\mathcal{E} \in \mathcal{C}(N,K)$ sei eine Matrix, deren Elemente nur Null oder Eins sind, und $e = (1,...,1)'$. Wir definieren $X = \mathcal{E} .\star m$. Dann gilt

a) $m = Me$
b) $X = M\mathcal{E}$
c) $X'e = \mathcal{E}'m$
d) $X'(a./m) = \mathcal{E}'a$ für beliebigen Spaltenvektor a.

Um eine Anpassung (Gewichtung, Redressment) der Häufigkeiten im ALLBUS an bekannte Randverteilungen für einige, für wichtig gehaltene Merkmale zu erhalten, fassen wir den ALLBUS als Stichprobe des Mikrozensus auf. Dies bedeutet keine Einschränkung, erleichtert aber die Darstellung der Idee. Wir ordnen die N Zellen des Mikrozensus lexikographisch an und bezeichnen mit

$$m = (m_1,...,m_N)' \quad \text{den Häufigkeitsvektor beim ALLBUS ,}$$

$$n = (n_1,...,n_N)' \quad \text{den Häufigkeitsvektor beim Mikrozensus}$$

für eine vorgegebene Merkmalsgruppe. Mit $\mathcal{E}$ soll die $N \times K$ Designmatrix bezeichnet werden, die die bekannten Randverteilungen $\mathcal{E}'n$ beim Mikrozensus liefert. Ausgangspunkt für eine Gewichtungsformel ist stets das tupel $(m, \mathcal{E}'n)$. Das interessierende Merkmal sei $\mathcal{Y}$.

$$\theta^A = (y_1^A,...,y_N^A)' \text{ ist der Häufigkeitsvektor von } \mathcal{Y} \text{ beim ALLBUS ,}$$

$$\theta^M = (y_1^M,...,y_N^M)' \text{ ist der Häufigkeitsvektor von } \mathcal{Y} \text{ beim Mikrozensus.}$$

Zu gegebenem tupel $(m, \mathcal{E}'n)$ ist ein Gewichtsvektor $a = (a_1,...,a_N)'$ gesucht, daß

$$(\sum_{i=1}^{N} a_i y_i^A - \sum_{i=1}^{N} y_i^M)^2$$

in einem zu präzisierenden Sinn möglichst klein wird. Außerdem ist Repräsentativität des Schätzers, das heißt

$$a'X = n'\mathcal{E} = (n./m)'X$$

gefordert, mit $X = \mathcal{E}. \star m$ und $.\star$ bzw. $./$ als elementweiser Multiplikation bzw. elementweiser Division.
Wir fassen θ^A und $e'(\theta^M - \theta^A)$ im Parametervektor $\theta = (y_1^A, ..., y_N^A, \sum_{i=1}^{N}(y_i^M - y_i^A))'$ zusammen, wobei $e = (1, ..., 1)'$ definiert ist. Wir haben $\sum_{i=1}^{N+1} \theta_i = \sum_{i=1}^{N} y_i^M$. Also ist $\sum_{i=1}^{N+1} \theta_i$ die zu schätzende Größe. Definiert man $s = \{1, ..., N\}$ und $r = \{N+1\}$, so gilt

$$(\sum_{i=1}^{N} a_i y_i^A - \sum_{i=1}^{N} y_i^M)^2 = (\sum_{i \in s} a_i \theta_i - \sum_{i=1}^{N+1} \theta_i)^2.$$

Wir wollen die vom Autor (1990b) eingeführte bedingte Minimax-Regel zur Bestimmung von a verwenden. Dazu sei $\tilde{V}$ eine symmetrische, positiv semidefinite $(N+1) \times (N+1)$ Matrix mit $\tilde{V}\tilde{X} = 0$. $\tilde{X}$ ist eine $(N+1) \times K$ Matrix, die sich aus der vertikalen Verkettung von X und dem Vektor $diff' = (n - m)'\mathcal{E}$ zusammensetzt. Die Aufgabe

$$\max_{\theta_{N+1}} \frac{(\sum_{i \in s} a_i \theta_i - \sum_{i=1}^{N+1} \theta_i)^2}{\theta' \tilde{V} \theta} \rightarrow Minimum$$

hat als Lösung

$$a = e_s - \tilde{V}_{sr} \tilde{V}_{rr}^{-1} e_r.$$

Dabei sind $\tilde{V}_{sr}$ bzw. $\tilde{V}_{rr}$ die entsprechenden Teilmatrizen von $\tilde{V}$ und $e_s = e = (1, ..., 1)'$ sowie $e_r = 1$. Es gilt $X'\tilde{V}_{sr}\tilde{V}_{rr}^{-1} = -diff$.

Da nach der Durchführung einer Erhebung die Daten bekannt sind, in unserem Fall θ^A, ist das Maximum nur bezüglich θ_{N+1} zu nehmen. Es hätte sich jedoch dieselbe Lösung ergeben, wenn wir das Maximum bezüglich ganz θ gebildet hätten. Das Maximum hat offensichtlich nur dann einen endlichen Wert, wenn der Zähler stets Null ist, im Falle der Nenner Null ist. Dies sichert die Repräsentativität.
Eine andere Interpretation der Minimaxaufgabe ist, diejenige Gewichtung a zu suchen, bei der der maximale Verlust $(\sum_{i \in s} a_i \theta_i - \sum_{i=1}^{N+1} \theta_i)^2$ minimiert wird, wobei der Parametervektor θ ein Element des Parameterraumes $\{\theta : \theta' \tilde{V} \theta \leq c^2\}$ ist. c ist eine beliebige Konstante.
Die Vorkenntnis über die interessierende Summe der Häufigkeiten in der Gesamtheit (hier: Mikrozensus) beschränkt sich daher auf das Wissen, daß sie innerhalb eines Intervalls liegt, dessen Grenzen von den $\mathcal{Y}$-Werten der Stichprobe (hier: ALLBUS) und von geschätzten Ausfallgrößen abhängen. Genau gilt:

$$\tilde{V}_{rr}[e'\theta^M - (e' - \tilde{V}_{rs}\tilde{V}_{rr}^{-1})\theta_A]^2 \leq c^2 - \theta_A'(\tilde{V}_{ss} - \tilde{V}_{sr}\tilde{V}_{rs}\tilde{V}_{rr}^{-1})\theta_A.$$

Wir haben

$$\theta_A'(\tilde{V}_{ss} - \tilde{V}_{sr}\tilde{V}_{rs}\tilde{V}_{rr}^{-1})\theta_A = \theta_A'(I \quad \vdots - \tilde{V}_{sr}\tilde{V}_{rr}^{-1})\tilde{V}\begin{pmatrix} I \\ \cdots \\ -\tilde{V}_{rs}\tilde{V}_{rr}^{-1} \end{pmatrix}\theta_A.$$

Wegen der positiven Semidefinitheit von $\tilde{V}$ ist dieser Ausdruck stets nichtnegativ.

Wir definieren

$$\begin{aligned} H &= X(X'X)^{-1}X' \\ u &= X(X'X)^{-1}diff = H(n./m - e) \\ \tilde{H} &= \tilde{X}(\tilde{X}'\tilde{X})^{-1}\tilde{X}' \end{aligned}$$

und erhalten mit $Hu = u$

$$\tilde{H} = \begin{pmatrix} H - \dfrac{uu'}{1+u'u} & \dfrac{u}{1+u'u} \\ \dfrac{u'}{1+u'u} & \dfrac{u'u}{1+u'u} \end{pmatrix}$$

$\tilde{V}$ können wir beispielsweise mittels einer Diagonalmatrix $\tilde{D} = diag(\tilde{d})$ mit $\tilde{d} = (d', \delta)'$ definieren durch

$$\tilde{V} = (I - \tilde{H})\tilde{D}(I - \tilde{H})$$

mit I als Identitätsmatrix passender Ordnung. Einige Umformungen ergeben

$$\tilde{V} = \begin{pmatrix} (I - H + \frac{uu'}{1+u'u})D(I - H + \frac{uu'}{1+u'u}) + \frac{\delta uu'}{(1+u'u)^2} & -(I - H + \frac{uu'}{1+u'u})D\frac{u}{1+u'u} - \frac{\delta u}{(1+u'u)^2} \\ -\frac{u'}{1+u'u}D(I - H + \frac{uu'}{1+u'u}) - \frac{\delta u'}{(1+u'u)^2} & \frac{u'Du+\delta}{(1+u'u)^2} \end{pmatrix}$$

Offensichtlich ist $\tilde{V}\tilde{X} = 0$. Formt man

$$a = e_s - \tilde{V}_{sr}\tilde{V}_{rr}^{-1}e_r$$

um, so erhält man mit $e = e_s$

$$\boxed{a = e + [I \quad + \quad \frac{1+u'u}{u'Du+\delta}(I - H)D]u}$$

Setzen wir $\delta = 1 + u'u - u'Du$, so ist

$$\boxed{a = e + [I \quad + \quad (I - H)D]u.}$$

Da sich jede Lösung der Gleichung

$$X'a = X'(n./m)$$

im Falle ihrer Existenz in der Form

$$a = X(X'X)^{-1}X'(n./m) + (I - X(X'X)^{-1}X')z$$

schreiben läßt, wobei z ein beliebiger $N-$dimensionaler Vektor ist, erhält man

$$\begin{aligned} a &= H(n./m) + (I - H)z \\ &= u + He + (I - H)z \\ &= e + [u + (I - H)(z - e)]. \end{aligned}$$

Definiert man $d = (z - e)./u$, ergibt sich mit $D = diag(d)$

$$a = e + [I + (I - H)D]u.$$

Mit $q = n./m$ ist

$$a = q - (I - H)(I - DH)(q - e).$$

Im Falle $D = I$, erhalten wir

$$a = e + u = q - (I - H)(q - e)$$

als Gewichte.

Die Matrix D braucht keine Diagonalmatrix zu sein. Es wird aber im folgenden Satz konstruktiv gezeigt, daß zu beliebig vorgegebenem tupel $(m, \mathcal{E}'n)$ stets ein $D = diag(d)$ existiert, so daß $a. \star m = n$ ist. Dies bedeutet jedoch nicht, daß bei fest vorgegebenen tupel $(m, \mathcal{E}'n)$, und damit festem u, auch alle Vektoren ν mit $\mathcal{E}'\nu \neq \mathcal{E}'n$ erreicht werden können, so daß $a. \star m = \nu$ wäre.

Satz. Gegeben sei das tupel $(m, \mathcal{E}'n)$. Erzeugt der Vektor m eine Matrix $X = \mathcal{E}. \star m$ vollen Ranges, dann existiert eine positiv definite Diagonalmatrix $D = diag(d)$, so daß $a. \star m = n$.

Beweis. Wir definieren

$$d = (n - m - u. \star m)./(u. \star m)$$

Für Null-Komponenten von $u. \star m$ können die entsprechenden d Komponenten beliebig gewählt werden. Wegen

$$\begin{aligned} HDu &= H.\star u.\star d \\ &= H.\star (n-m-u.\star m)./m \\ &= X(X'X)^{-1}X'(n-m-u.\star m)./m \\ &= X(X'X)^{-1}\mathcal{E}'(n-m)-u \\ &= 0 \end{aligned}$$

ist

$$\begin{aligned} a.\star m &= [e+(I+(I-H)D)u].\star m \\ &= m+u.\star m+[(I-H)Du].\star m \\ &= m+u.\star m+(Du).\star m \\ &= m+u.\star m+u.\star d.\star m \\ &= m+u.\star m+n-m-u.\star m \\ &= n \end{aligned}$$

Offensichtlich ist $d=(n-m-u.\star m)./(m.\star u)$ nicht notwendigerweise positiv. Wir wählen daher γ so, daß $d+\gamma e$ positiv ist. Weil $Hu=u$ ist, gilt $(I-H)(D+\gamma I)u=(I-H)Du$; daher bleibt a unverändert.

◇

Unabhängigkeit. Wie zuvor gezeigt wurde, ist es möglich, zu vorgegebener Randverteilung alle zulässigen Häufigkeitsvektoren über die Gewichtung a zu erzeugen. Es befindet sich daher auch die Gewichtung darunter, die man bei der Verwendung des IPF-Algorithmus erhalten würde. Wir wollen es an einem Beispiel zeigen.

Beispiel : 2 × 2 Tabelle.

Gegeben sei die Ausgangstabelle

1	1	2
1	1	2
2	2	4

Gesucht ist eine Tabelle mit folgenden Randhäufigkeiten

		3
		7
4	6	10

Für

$$\mathcal{E} = \begin{pmatrix} 1 & 0 & 0 \\ 1 & 0 & 1 \\ 1 & 1 & 0 \\ 1 & 1 & 1 \end{pmatrix} = X \; ; \; H = \tfrac{1}{4} \begin{pmatrix} 3 & 1 & 1 & -1 \\ 1 & 3 & -1 & 1 \\ 1 & -1 & 3 & 1 \\ -1 & 1 & 1 & 3 \end{pmatrix}$$

setzen wir $\delta = 19,25$ und $D = diag(1,2,2,3)$. Wir erhalten $u = (0,1,2,3)'$ und daher

$$a = e + [I \quad + \quad \frac{1+u'u}{u'Du+\delta}(I-H)D]u = (6,9,14,21)'/5.$$

Dies wäre auch die IPF-Lösung.

Daß es viele Möglichkeiten gibt, D mit dieser Lösung für a zu wählen, zeigt sich, wenn wir $I - H = bb'$ schreiben mit $b = (-1,1,1,-1)'/2$. Dann ist

$$a = e + u \quad + \quad \frac{1+u'u}{u'Du+\delta} b(b'Du)$$

und alle Paare (D,δ) mit $\frac{1+u'u}{u'Du+\delta} b'Du = konstant$ führen zur selben Lösung.

Im Falle der Anpassung an nur eine Variable ist $X = I$. Daher ergibt sich für die Minimax-Gewichtung $a = e + u = n./m$ und damit Identität mit der IPF-Lösung.

Welche Eigenschaften sollte eine Gewichtung haben?
1. Die Gewichte sollten positiv sein. Diese intuitiv zunächst einleuchtende Forderung ist für die Minimax-Gewichtung nicht bei jeder Spezifikation erfüllt. Für ungünstig konditionierte Stichproben können negative Hochrechnungsfaktoren auftreten (vgl. auch Merz 1983:345). Im Beispiel der Anpassung des ALLBUS an den Mikrozensus trat dieser Fall nicht auf. Ist man zudem nur an der Schätzung $\sum a_i y_i^A$ für $\sum y_i^M$ interessiert, kann die Schätzung trotz einiger negativer Gewichte positiv sein. Es bleibt noch zu bedenken, ob negative Gewichte nicht ein Indiz dafür sind, daß die Reparaturtechnik zu gewaltsam sein muß. Die Randverteilungen der Ursprungstabelle liegen zu weit weg von den gewünschten Randverteilungen. Vorsicht vor künstlichen Gewichtungen ist dann auf jeden Fall geboten.

2. Ändert man die Dimension der Ausgangstabelle, multipliziert man also die Starthäufigkeiten mit einer Zahl c, sollte die neue Gewichtung aus der alten durch Division mit c hervorgehen. In Formeln:

$$\begin{array}{lcll} m & \longrightarrow & m_c & = c \cdot m \\ X & \longrightarrow & X_c & = \mathcal{E} .* m_c = c \cdot X \\ H & \longrightarrow & H_c & = H \\ u & \longrightarrow & u_c & = [u + (1-c)He]/c \\ a & \longrightarrow & a_c & = [a + (1-c)(I-H)(DH-I)e]/c. \end{array}$$

Für $c \neq 1$ muß gelten: Es existiert ζ mit

$$(DH - I)e = H\zeta$$

das heißt

$$d = (e + H\zeta)./(He).$$

Lineares Modell.

Wir betrachten nun das lineare Modell $\tilde{Y} = \tilde{X}\beta + \epsilon$, wobei $\mathrm{E}(\epsilon) = 0$ und $\mathrm{E}(\epsilon\epsilon') = \tilde{\Sigma}$. $\tilde{\Sigma}$ ist eine positiv semidefinite Matrix definiert durch $\tilde{\Sigma} = (I - \tilde{X}\tilde{G}')\tilde{V}^{+}(I - \tilde{G}\tilde{X}')$. $\tilde{G}$ ist eine $(N+1) \times K$ Matrix mit $\tilde{G}'\tilde{X} = I$. A^{+} ist die Moore-Penrose-Inverse der Matrix A. Analog zu oben sei $Y = Y_s = (Y_1, ..., Y_N)'$ und $Y_r = Y_{N+1}$. Bezüglich des Untermodells Y lautet nach Rao/Mitra (1971 S.148) der BLU Schätzer $\hat{\beta}$ für β

$$\hat{\beta} = (X'(\Sigma + XX')^{-1}X)^{-1}X'(\Sigma + XX')^{-1}Y$$

Wie gezeigt werden kann (Gabler 1990a), läßt sich $\hat{\beta}$ auch in der Form

$$\hat{\beta} = (G_s' - G_r'\tilde{V}_{rr}^{-1}\tilde{V}_{rs})Y$$

schreiben, mit G_s als den ersten N Zeilen von $\tilde{G}$ und G_r als der letzten Zeile von $\tilde{G}$. Weiter gilt $cov(\tilde{G}'\tilde{Y}) = \tilde{G}'\tilde{\Sigma}\tilde{G} = 0$. Folglich ist $\tilde{G}'\tilde{Y} = E(\tilde{G}'\tilde{Y}) = \tilde{G}'X\beta = \beta$ mit Wahrscheinlichkeit Eins.

Wir definieren $\hat{\tilde{Y}} = \tilde{X}\hat{\beta}$ und erhalten

$$\sum_{i=1}^{N+1} \hat{Y}_i = e'X\hat{\beta} + (n-m)'\mathcal{E}\hat{\beta} = n'\mathcal{E}\hat{\beta} = n'\mathcal{E}(X'(\Sigma + XX')^{-1}X)^{-1}X'(\Sigma + XX')^{-1}Y.$$

Für $Y = m$ gilt offensichtlich $\hat{\beta} = (1, 0, ..., 0)'$ und daher

$$\sum_{i=1}^{N+1} \hat{Y}_i = n'\mathcal{E}\hat{\beta} = n'e = \sum_{i=1}^{N} n_i = e'(a. * m) = a'm.$$

Im allgemeinen ist jedoch $n'\mathcal{E}(X'(\Sigma + XX')^{-1}X)^{-1}X'(\Sigma + XX')^{-1} \neq a'$, wobei $a = e_s - \tilde{V}_{sr}\tilde{V}_{rr}^{-1}e_r = e + [I + \frac{1+u'u}{u'Du+\delta}(I - H)D]u$ die Minimax-Gewichtung ist.