大阪大学教育学年報 第5号（2000年）
Annals of Educational Studies Vol .5

欠損データの分析がもたらす新たな知見
—1995年SSM調査の学歴移動表分析を例として—

保 田 時 男

【要旨】
調査項目への無回答などにより発生する欠損データは、しばしば分析結果の歪曲や精度の低下といった問題の源として警戒される。しかし、欠損データを適切に分析すれば、その分析結果はむしろ豊富な情報を提供してくれ、データの新たな解釈の可能性を与えてくれる。本稿は、1995年SSM調査による学歴移動表の分析を例として、欠損データ分析の意義について解説している。本稿で用いたログリニア・モデルの応用による分析方法は、欠損データによるデータの偏りが「無視できない」場合を含めた幅広いモデルを対比するのに有効である。分析対象の学歴移動表は、父親の学歴から本人の学歴への世代間移動を検討しているものだが、父親の学歴についての無回答が多く存在する。この欠損データを分析した結果、新制教育を受けた年齢層の者が欠損率を高めているモデルと、父親の学歴が高い者が欠損率を高めているモデルという2つの可能性が明らかになった。この結果は父学歴無回答の原因に関する知見を与えてくれ、学歴移動表の解釈幅を拡大してくれる。特に後者のモデルの場合、学歴移動の開放性が従来よりも大きく見直されることになり、欠損データの分析により得られる情報の重要性が示されている。

1. 分析の目的

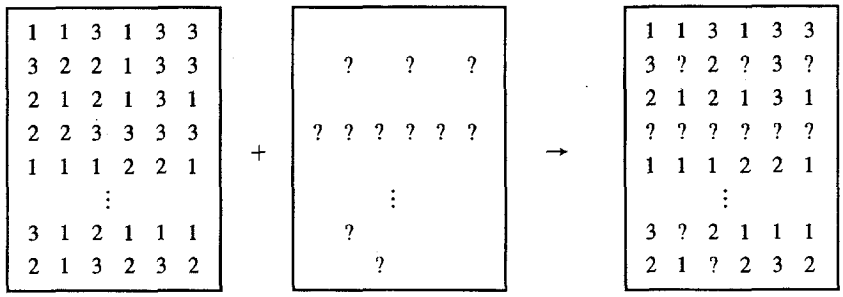
1-1. 欠損データの根本的な問題性

父親の学歴から子どもの学歴への世代間学歴移動について社会調査データを分析しようとすると、父学歴について無回答のケースが多く、分析の障害になる。本稿は、この無回答による欠損データ（missing data）に注目し、ログリニア・モデルを応用した方法でその影響と原因の分析を試みている。欠損データが計量分析に与える問題性は大きいが、本稿の目的は欠損データ問題を回避するという消極的な理由にあるのではなく、むしろ欠損データを分析の視野に含めることで分析の幅が広がる可能性を積極的に示すことにある。学歴移動の分析はその例示である。

とは言え、欠損データを分析するためにはその問題性を理解する必要があるので、やや遠回りになるが、まずその根本的な問題性について確認しておく。欠損データとは、「本来与えられるべきはっきりとした値が存在すると想定できるが、実際には取得することができなかったデータ部分の総称」である。社会調査によるデータ取得の過程を表した模式図（図1）で説明してみよう。調査対象となるサンプルが持つデータは本来［1－a］のように完全な表として表される。しかし、これらのデータの全てが調査によって首尾よく回収されるとは限らず、何らかの原因で値が調査者の手に渡らないことがある。その原因は、調査対象者が回答を秘密にしたことによる場合もあれば、本当に知らないことによる場合もある。調査対象

者の不在や協力拒否によってすべての変数についてデータが得られないこと（unit nonresponse）もあれば、一部の変数が無回答であり部分的にデータが得られないこと（item nonresponse）もある。

図1 欠損データの模式図



[1-a] [1-b] [1-c]

ともかく、何らかの原因によってデータが得られなかった部分はちょうど［1－b］の「？」シールが上から張られたように隠されてしまい、実際の調査データは一部のデータが不明な表［1－c］となる。［1－c］の「？」部分が欠損データである。

欠損データの定義においては、本来与えられるべき値（［1－a］）が存在することが重要な点であり、本来与えられるべき値がはっきりしない場合は、それを欠損データとは呼ばない。このため、データ分析上欠損値（missing value）を当てられるものの中にも欠損データとはみなせないものも存在する[1]。Little & Rubin（1987, pp.3）は欠損データとみなせない無回答の例として政策の選好を尋ねる世論調査への無回答を挙げている。選好の無回答の裏にはっきりとした選好が隠されているかどうかは定かではなく、全く政策に関心がないのかもしれない。この無回答は欠損データとみなすよりは独自のカテゴリとみなす方が適切と言える。

欠損データが発生した場合、よく取られる対処方法は、欠損データを含むケースを取り除いて分析を行うことであり、この方法はリストごとの除去（listwise deletion）と呼ばれる。しかし、リストごとの除去は欠損データがもたらしうる2つの問題性の両方ともに抵触する。第1の問題は、回答が得られているケースと無回答のケースでデータ内容に差異がある可能性により発生する。このとき、リストごとの除去は目的とする母集団の傾向を正しく代表せず、回答ケースの偏った特徴を表現してしまう。これを「偏りの問題」と呼ぶことにしよう。もう1つの問題は、除去したケースに含まれている欠損していない変数についての情報をも捨ててしまうことに起因する。得られている情報を完全には用いていないので、分析結果の精度（つまり信頼性）が正しく評価されない。これを「精度の問題」と呼ぶことにしよう。

1-2.「無視できる」欠損メカニズムと「無視できない」欠損メカニズム

計量社会学者にとって重要な問題は「精度の問題」というよりは「偏りの問題」の方である。計量分析の第一の目的はデータを要約したモデルを導き出すことであるが、「偏りの問題」はモデルの選定に大きな影響を及ぼすことがあるからである。一方の「精度の問題」はマーケティング調査の信頼性や投薬の効果の程度を正確に測定する必要がある経営や医療の分野で問題にされることが多い。

そのため、「偏りの問題」の有無と内容を判別することは欠損データを分析する計量社会学者の重要な課題になるが、データが欠損しているわけであるからそれを完全に判別することはできない。そこで欠損データの分析者が取る戦略は、欠損が発生する原因メカニズムをさまざまに仮定し、その中でどのような「偏りの問題」を引き起こす仮定がデータに適合するかを検定するという方法である。適合する仮定はほとんどの場合複数あり、許容される仮定の幅は欠損データが本来持つ不確かさの幅を表現していることになる。もし、重大な「偏りの問題」を引き起こす仮定が受け入れられる様であれば、その偏りの可能性を考慮したデータ解釈をしなければならないと言える。

「偏りの問題」の有無は分析者が仮定する欠損メカニズム（missing mechanism, missing-data mechanism）に完全に依存する。「偏りの問題」を引き起こす仮定を「無視できない（nonignorable）」欠損メカニズム、「偏りの問題」を含まない仮定を「無視できる（ignorable）」欠損メカニズムと呼び（Little & Rubin 1987, pp.8-13）、ある欠損メカニズムが「無視できる」ための必要十分条件は、そのメカニズムがMAR（missing at random）であることとすでに証明されている（Rubin 1976）。MARとは欠損の確率が観察されているデータにのみ依存し、観察されていないデータからは独立であるという意味であり、式で書くと式（1）のようになる（Schafer 1997, pp.11）。ただし、Rは各ケースの各変数が欠損しているかどうかを示す行列（図1の［1−b］）であり、Y_{obs}、Y_{mis}はそれぞれ調べられているサンプルY（図1の［1−a］）のうち観察できた部分と観察できなかった部分である。ξは未知のパラメータによる誤差を表す[2]。具体的には、父学歴の無回答発生率が隠されている父学歴の内容には依存せずに、その他の観察されている情報（年齢や子どもの学歴など）にのみ依存するということである。

$$\Pr(R \mid Y_{obs}, Y_{mis}, \xi) = \Pr(R \mid Y_{obs}, \xi) \tag{1}$$

MARは「偏りの問題」が発生しないことを保証するだけであるが、さらにサンプル数の減少という以外に「精度の問題」を起こさない欠損メカニズムも存在する。そのような欠損メカニズムはMCAR（missing completely at random）と呼ばれ、ある変数が欠損する確率が全くデータに依存しないことを意味する。MCARを式で表すならば、式（2）のように表される。

$$\Pr(R \mid Y_{obs}, Y_{mis}, \xi) = \Pr(R \mid \xi) \tag{2}$$

リストごとの除去という一般的な方法は、欠損メカニズムにMCARを仮定することと同義

である。MCARは非常に限定の強い欠損メカニズムであり、リストごとの除去は広範な欠損メカニズムの中の狭い1点を仮定することで分析を進めていると言える。MCARに限定することなくさまざまな欠損メカニズムを仮定した分析をおこなえば、データが本来持つ解釈可能性の幅を適切に捉えることができる。そこから未知なる知見が得られる可能性が広がる所に欠損データの分析を行う積極的な価値が存在するのである。

1-3. ログリニア・モデルによる欠損データ分析の位置付け

欠損データを分析する具体的な方法には実にさまざまのものがあるが、本稿ではChen & Fienberg（1974）に端を発し、Fay（1986）やBaker & Laird（1988）により一般化されたログリニア・モデルを応用した方法を用いている。これはこの方法が計量社会学者の一般的な関心を非常によく満たしてくれるからである。

欠損データの分析方法として現在もっとも展開が激しいのは、間違いなく多重帰属推定（multiple imputation）と呼ばれる別の方法であるが、この方法は欠損メカニズムを「無視できる」MARに限定することで、発展している。多重帰属推定の焦点は欠損メカニズムの違いによる「偏りの問題」にではなく、「無視できる」欠損メカニズムを仮定することによる信頼性の低下を正確に測定すること、つまり「精度の問題」に向けられているのである[3]。

一方、ログリニア・モデルによる欠損データ分析はその発展過程の中で、「無視できない」欠損メカニズムを適切に表現する方法を確立しており、その方向に積極的に利用されている。ログリニア・モデルを応用しているのでカテゴリ欠損データしか扱えない[4]という制限を持つものの、社会調査により得られるデータの多くがカテゴリカルであることを考えると、それは我々にとってあまり大きな制約にはならないであろう[5]。

当初、Chen & Fienberg（1974）が用いた方法は1）2変数のクロス集計表で、2）「無視できる」欠損メカニズムにモデルの範囲が限られていたが、これらの限定は徐々に解消され、最終的には、1）多変量のクロス集計表で、2）「無視できない」欠損メカニズムを含んだモデルが扱えるものに一般化されている[6]。「無視できる」欠損メカニズムを仮定することができ、社会学者にとってなじみのあるログリニア・モデルの応用で分析を行うことができるこの方法は、社会学者の関心を満たすための有効な道具となり、本稿における学歴移動表の分析に対しても適切と言える。

2. 分析対象と方法

2-1. 分析対象の学歴移動表

本稿で分析の対象としているのは、1995年SSM調査の結果から作成した父親の学歴から本人の学歴への学歴移動表（表1）である。世代間の学歴移動は、教育機会の平等性、開放性を判断する材料としてしばしば教育社会学で問題にされるが、一般に親の学歴への無回答が多いという特徴を持つ。この無回答は本来与えられるべき何らかの回答を必ず持っているので、欠損データと言える。その無回答率は近年高まっており[7]、欠損の原因や欠損による分

析結果の偏りを考察することは、学歴移動表をより適切に解釈する上でますます重要になっている。

分析は本人の性別で別個に行い、それぞれの性別について本人の年齢層（*A*）、父親の学歴（*B*）、本人の学歴（*C*）の3変数によるクロス集計表を分析している。年齢層（*A*）は調査対象者本人の年齢を10年ごとに区切っている。70歳の対象が少数いるが60代にまとめた。父学歴（*B*）は対象者の父親の最終学歴を、新制・旧制の区別なく高等・中等・義務の3つに分けたものである。旧制高等小学校を義務に含めたことを除いては文字通りの分類である（表1の注を参照）。本人学歴（*C*）の分類も同様であり、調査対象者本人の最終学歴を表している。なお、本人学歴について無回答を示しているケースが微量存在するが、このケースは分析対象から除外している。

表1　1995年SSM調査の性別・年齢層別学歴移動表

		男性					女性				
		C: 本人学歴					C: 本人学歴				
A: 年齢層	B: 父学歴	高等	中等	義務	無回答	計	高等	中等	義務	無回答	計
20代	高等	52	11	1	0	64	66	20	2	0	88
	中等	58	70	3	1	132	65	86	6	0	157
	義務	24	68	9	0	101	21	80	8	0	109
	無回答	14	31	6	0	51	9	23	5	2	39
	計	148	180	19	1	348	161	209	21	2	393
30代	高等	58	18	0	0	76	72	32	0	0	104
	中等	46	55	1	0	102	82	81	0	0	163
	義務	56	105	15	0	176	39	143	9	0	191
	無回答	18	32	6	0	56	23	59	3	1	86
	計	178	210	22	0	410	216	315	12	1	544
40代	高等	60	19	1	0	80	37	41	4	0	82
	中等	59	40	8	0	107	50	95	11	0	156
	義務	80	219	61	0	360	40	256	65	0	361
	無回答	10	69	40	0	119	14	98	43	0	155
	計	209	347	110	0	666	141	490	123	0	754
50代	高等	31	10	5	0	46	18	27	4	0	49
	中等	22	34	18	0	74	18	52	18	1	89
	義務	32	154	104	1	291	11	160	143	2	316
	無回答	9	43	50	0	102	4	77	90	1	172
	計	94	241	177	1	513	51	316	255	4	626
60代	高等	34	5	3	0	42	5	12	4	0	21
	中等	17	25	10	0	52	9	47	8	0	64
	義務	49	127	197	0	373	13	128	210	0	351
	無回答	10	22	53	1	86	2	30	82	0	114
	計	110	179	263	1	553	29	217	304	0	550

注：高等＝新制短大以上・旧制高校以上
中等＝新制高校・旧制中学以上
義務＝新制中学・旧制高等小学校以下

2-2. 分析方法の概略

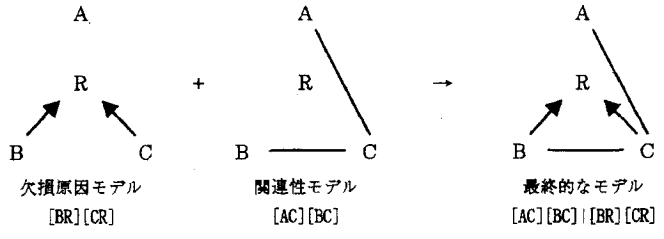
本稿で用いたログリニア・モデルを応用した欠損データの分析方法について簡単に説明する。より詳しくはFay（1986）、Baker & Laird（1988）、保田（2000）などを参照してほしい。分析を行うためには、まず元のクロス集計表を図1に表した欠損データ発生の仕組みに沿うように組み立てなおさなければならない。我々が手にしているデータ［1－c］は、本来得られるべき値の表［1－a］とそれぞれの変数について欠損しているかどうかを表す表［1－b］が掛け合わされた結果である。ここでは、［1－a］は5×3×3のクロス集計表であり、［1－b］は2値のダミー変数（*R*）で表される。このダミー変数はしばしば欠損指標（missing indicator, missing-data indicator）と呼ばれる（Little & Rubin 1987, pp.89）。［1－a］と［1－b］を掛け合わせた(5×3×3)×2のクロス集計表が分析の対象になる。この表の各セル度数を式（3）のように表すことにする。

$$f_{abcr}\left(a=\begin{cases}1=20\text{代}\\ \vdots\\ 5=60\text{代}\end{cases},b=\begin{cases}1=\text{高等}\\ 2=\text{中等}\\ 3=\text{義務}\end{cases},c=\begin{cases}1=\text{高等}\\ 2=\text{中等}\\ 3=\text{義務}\end{cases};r\equiv\begin{cases}1=\text{父学歴が欠損でない}\\ 2=\text{父学歴が欠損している}\end{cases}\right) \tag{3}$$

当然、f_{abc2}の値は父学歴（*B*）が欠損しているので不明であり、データを集計するだけではこの架空の表を作成することはできない。検証したいモデルの仮定に基づいて最尤推定値$\hat{f}_{abcr}$を推定しなければならない。

検証されるモデルは、通常のログリニア・モデルにおける変数間の関連性（独立性・非独立性）を表現するモデルと欠損メカニズムを表現するモデルの合成モデルとして組み立てられる。それぞれ仮に関連性モデル、欠損原因モデルと呼ぶことにしよう。例えば、欠損原因モデルが[BR][CR]で、関連性モデルが[AC][BC]の場合、モデルは最終的に図2のように合成される。モデルに基づいた推定値$\hat{f}_{abcr}$はEMアルゴリズム（Dempster et al. 1977、MacLachlan & Krishnan 1997）と呼ばれる最尤推定法の一種により導き出される。EMアルゴリズムはモデルで想定されている変数間の因果、関連の限定の中で最もデータに適合的な推定値を導き出す。

図2 モデルの合成



そうして導き出された推定値が実測データをよく再現できているかどうかは、尤度比統計量（Likelihood ratio statistic）による検定という標準的な方法を用いればよい。ただし、比較する度数はやや変則的で、欠損していない各セル f_{abc1} および欠損しているセルの（この場合は）Bについての各マージン f_{ac2}（$=\sum_b f_{abc2}$）を対比させる。尤度比統計量（L^2）は式（4）により算出され、モデルによる推定度数の実測値に対する当てはまりが悪いほどこの値が大きくなる。

$$L^2 = 2\sum f_{abc1} \bullet \log\left(f_{abc1} / \hat{f}_{abc1}\right) + 2\sum f_{a\bullet c2} \bullet \log\left(f_{a\bullet c2} / \hat{f}_{a\bullet c2}\right) \qquad (4)$$

3. 分析結果

3-1. モデルの選定

表1の学歴移動表に対して実際にさまざまな欠損原因モデルを適用してみたところ、各モデルの適合度検定の結果は表2のようになった。関連性モデルは変数間の全関連性を認める飽和モデルでしか適合しなかったが、欠損原因モデルはかなり単純なモデルである[AR][CR]モデルや[BR][CR]モデルで適合している。ただし、女性の[BR][CR]モデルについては、1段階複雑な[AR][BR][CR]モデルとの間の適合度の差が大きすぎるので、採択できない。結局、男性は[ABC]｜[AR][CR]と[ABC]｜[BR][CR]の2種類、女性は[ABC]｜[AR][CR]のみがデータを的確に表現しているモデルとみなされる。

表2　各モデルの適合度検定の結果

欠損原因モデル		男性 L^2	女性 L^2
[R]	df=14	101.30	138.87
[AR]	df=10	92.44	79.62
[BR]	df=12	46.26	32.56
[CR]	df=12	22.23	29.40
[AR][BR]	df=8	30.59	17.83
[AR][CR]	df=8	11.34 *	2.86 *
[BR][CR]	df=10	10.45 *	16.45 *
[AR][BR][CR]	df=6	8.83 *	1.79 *

注：関連性モデルは飽和モデル[ABC]
*は5%水準で適合。網がけは採択されるモデル

先に述べた通り、欠損データによる「偏りの問題」が発生するかどうかは、仮定した欠損メカニズムが「無視できる」か「無視できない」かによる。採択された欠損メカニズムの中で[BR][CR]モデルは「無視できない」が、[AR][CR]モデルは[BR]を含まないMARなので「無視できる」。したがって、「無視できる」欠損メカニズムのみが採択されている女性の学歴移動表については「偏りの問題」の可能性が否定されたことになる。「偏りの問題」の可能性がある方がより興味深いので、以降は男性に対象を絞った分析結果を表すことにする。

3-2. モデルがデータ解釈に与える影響

男性の[AR][CR]モデル、[BR][CR]モデルそれぞれについて、欠損に隠された父学歴を最尤推定した結果が表3である。[AR][CR]モデルではほとんどのケースの父学歴が義務（あるいは若い層では中等）に配分されているのに対して、[BR][CR]モデルでは義務と高等に二分化している。この両方の配分がデータに適合していることから欠損データがもたらす不確かさがいかに大きいかが分かる。

表3 男性の各モデルにおける欠損データの最尤推定値

		[AR][CR]モデル			[BR][CR]モデル		
		C: 本人学歴			C: 本人学歴		
A: 年齢層	B: 父学歴	高等	中等	義務	高等	中等	義務
20代	高等	5.0	2.4	0.4	13.4	27.9	5.1
	中等	5.6	15.2	1.3	0.0	0.0	0.0
	義務	2.3	14.8	4.0	0.1	3.4	1.5
30代	高等	5.2	3.6	0.0	15.4	33.0	2.8
	中等	4.2	10.9	0.4	0.0	0.0	0.0
	義務	5.1	20.8	5.9	0.3	5.2	2.5
40代	高等	5.8	4.5	0.5	14.1	55.4	26.8
	中等	5.7	9.4	3.9	0.0	0.0	0.0
	義務	7.7	51.5	30.0	0.4	10.9	10.3
50代	高等	2.9	2.1	2.0	8.1	32.3	34.4
	中等	2.1	7.2	7.4	0.0	0.0	0.0
	義務	3.0	32.7	42.5	0.2	7.7	17.2
60代	高等	2.2	0.7	0.8	8.9	14.7	21.6
	中等	1.1	3.5	2.7	0.0	0.0	0.0
	義務	3.1	17.8	53.2	0.2	6.3	32.7

この2つの配分の違いが実際の学歴移動研究に及ぼす影響力を見るための指標として表4の安田の開放性係数を算出した。安田（1971, pp.59-61）によれば、社会移動は全体的な底上げなどの構造変化による強制移動と、移動への障壁低下による純粋移動の2種類に分けられる。このとき、純粋移動による移動の割合が高いことは、その社会の開放性、平等性を表していると言える。純粋移動率を一般的に比較するための指標として、安田は開放性係数（＝観測された純粋移動率／完全移動のときに期待される純粋移動率）を提案しており、学歴移動の研究においてこの係数はしばしば利用されている。

表4 男性の各モデルにおける安田の開放性係数

	リストごとに除去			[AR][CR]			[BR][CR]		
	強制移動率	純粋移動率	開放性係数	強制移動率	純粋移動率	開放性係数	強制移動率	純粋移動率	開放性係数
20代	0.30	0.26	0.71	0.30	0.26	0.70	0.25	0.33	0.82
30代	0.45	0.19	0.66	0.45	0.18	0.64	0.39	0.24	0.76
40代	0.53	0.18	0.75	0.52	0.17	0.73	0.41	0.32	0.97
50代	0.40	0.19	0.70	0.37	0.20	0.71	0.33	0.29	0.84
60代	0.33	0.10	0.44	0.33	0.10	0.44	0.27	0.19	0.61

参考のため一番左にリストごとの除去で欠損データを無視したときの開放性係数を示しているが、[AR][CR]モデルによる結果はこれとほぼ変わらない。この結果は[AR][CR]モデルが「無視できる」欠損メカニズムであることを改めて示している。一方、「無視できない」欠損メカニズムである[BR][CR]モデルにおける開放性係数は、どの年齢層でも他の2つのモデルに比べて高い値を示している。つまり、父学歴の内容が欠損率に影響を与えているとすれば、学歴移動の開放性は格段に上方修正されるのである。

3-3. 欠損メカニズムの詳細

表5 男性の各モデルにおける欠損原因パラメータ

		[AR][CR]		[BR][CR]	
		λ	exp(λ)	λ	exp(λ)
λ_r	R=2	-0.84	0.43	-2.22	0.11
λ_{ar}	(A,R)=(1,2)	0.05	1.05		
	(A,R)=(2,2)	0.01	1.01		
	(A,R)=(3,2)	0.08	1.08		
	(A,R)=(4,2)	0.03	1.03		
	(A,R)=(5,2)	-0.17	0.84		
λ_{br}	(B,R)=(1,2)			2.51	12.28
	(B,R)=(2,2)			-3.04	0.05
	(B,R)=(3,2)			0.53	1.70
λ_{cr}	(C,R)=(1,2)	-0.39	0.68	-0.98	0.38
	(C,R)=(2,2)	0.03	1.03	0.19	1.21
	(C,R)=(3,2)	0.36	1.44	0.79	2.21

分析結果の差異がどのような欠損メカニズムによってもたらされているのかをより詳細に検討するために、それぞれのモデルの欠損原因を表すパラメータを表5に示した。パラメータの意味は通常のログリニア・モデルと同様であり、例えば$\lambda_{ar}(1,2)=0.05$は年齢層が20代（$A=1$）のとき父学歴が欠損（$R=2$）の確率が平均のexp(0.05)＝1.05倍であることを意味する。

本人学歴が欠損率を左右する原因になっていること（[CR]）は両方のモデルに共通であるが、その影響の傾向もほぼ共通で、本人学歴が低いことが欠損率を高める原因になっている。ただし、その影響力は[BR][CR]モデルの方でより極端であり、本人学歴が義務であると平均の2.21倍も欠損率が高まることを示している。

[AR][CR]モデルにおけるもう1つの欠損原因である年齢層は、60代で欠損率をやや低める働きを示すがより若い年齢層における差異はほとんどない。50代と60代の間に欠損率を高める原因となる何らかの変化があったと考えることができる。この区切りはほぼ戦後の新制教育を受けた者とそうでないものの年齢区分に対応している。学制の転換により格段に多くの人々が中等以上の学歴を有するようになったので、「親の学歴は当然義務」との確信が弱まり、欠損率を高めたと解釈することができる。

一方、[BR][CR]モデルにおいて欠損原因となっている父学歴については、父学歴が高いことが欠損率を非常に高めることを示している。この結果は父学歴欠損ケースに高学歴からの

下降移動者が多いことを意味しており、これにより安田の開放性係数が大きくなっていたことが分かる。このメカニズムは保田（2000）による1985年SSM調査に対する分析結果とほぼ共通している。下降移動者が父学歴を無回答にする理由は、学歴に無関心な子どもが親の高学歴をあまり把握していないことによるという解釈が可能である。あるいは、下降移動を恥じての秘密とも捉えられるが、本人の学歴を回答していることからこれはあまり考えにくい。

4. 結論と議論

4-1. 分析結果から広がる学歴移動表の解釈

以上の結果をまとめると、次のようになる。学歴移動表について欠損データを含めた分析を行った結果、女性については本人学歴と年齢層が欠損率を左右するという「無視できる」モデルしか許容されなかったが、男性については、それとは別に本人学歴と父学歴そのものが欠損率に影響を与える「無視できない」モデルがデータに適合した。男性データに注目した分析を進めると、「無視できない」欠損メカニズムによるモデルでは、学歴の下降移動の量が大きく見直されるために安田の開放性係数が通常よりも大きくなることが分かった。

これらの結果から欠損データの分析がデータ解釈の新たな可能性を広げたことは明らかであるが、はたして拡大した解釈幅の中で結局どの解釈が選択されるべきなのかという問題が残される。この問題は、欠損データの分析結果からは単純に解決できないので、他のデータや既知の知見から総合的に判断するより他にない。ただ、今回の分析に関する限り、[AR][CR]モデルを採用する方が無難なように思える。その理由は、第1に男性と女性の両方について同じモデルで説明ができ、簡約的であること、第2にこのモデルにおける欠損メカニズムと他方のモデルにおける欠損メカニズムを対比すると、[AR][CR]モデルの方がより現実的に感じられること、が挙げられる。[AR][CR]モデルにおいては学制の転換時期に非常によく対応した欠損メカニズムが見られたのに対して、[BR][CR]モデルにおける父学歴高等の極端な影響は、やや不自然である。

[AR][CR]モデルは「無視できる」欠損メカニズムであり、こちらが採用されれば、欠損データによる「偏りの問題」は起こらないことになるが、それは欠損データ分析の意味をなくすものではない。父学歴の無回答原因に関する分析結果はやはり新たな情報として興味深いと言えるし、モデルが限定され欠損の不確かさが減少させられたことにも意味がある。結果に関わらず、欠損データの分析はデータの解釈に新たな知見を加えてくれるのである。

4-2. 誤答の考慮による新たな可能性

最後に、欠損データ分析の発展可能性の一方向として、欠損データを考慮した枠組み（図1）と同じ発想で、誤答の問題が扱えることについて述べておきたい。図1において真の値（[1−a]）の上にかぶせられるシール（[1−b]）には、データを隠す「？」しか描かれていないが、理念的にはここに誤った回答を記し別の値にみせかけることも可能である。欠損と誤答の可能性を同時に考えることで、データ解釈の幅はさらに広がる。本稿の学歴移動の分

析についても本人学歴の「嘘」を考慮に入れれば、また変わった結果が現れる可能性がある。

また、一般に「秘密」や「嘘」といった一連の欺瞞に関する研究領域に貢献できると考えられる。「秘密」や「嘘」に関しては社会学や倫理学の視点からいくらかの研究がなされているが、その研究結果を実証的に検証することは難しい。しかし、欠損データと誤答の可能性を視野に入れた計量分析がそれを可能にしてくれるかもしれない。

誤答の可能性を考えることは、連続量の変数についても可能であるが、回答のパターンが限定されているという意味からカテゴリカル・データの方が扱いやすいであろう。実際に本稿で用いたのと似た方法を使って、誤答（misclassification）によるクロス集計表の偏りについていくつかの研究がなされており（Press 1968、Tenenbein 1972、Chen 1979、Nordheim 1984）、新たな分析領域の開拓が期待される。

欠損データを含めた計量分析には、いまだ未知の部分や問題点が多い。しかし、本稿で明らかにされたように、欠損データの分析がもたらしてくれる新しい情報の価値は高い。特にログリニア・モデルを用いた欠損データ分析は、社会学者の関心を満たすための有効な手段としてさまざまな分野で利用可能であろう。欠損データの分析は時に解釈困難な分析結果の謎を解いてくれる鍵になる。単純に分析の対象から除外してしまうのではなく、未知の可能性を探るべきである。

【付記】
データの利用については1995SSM研究会の許可を得た。

＜注＞

1）社会調査における欠損値はDK（Don't Know）、NA（No Answer）等と表記されることが多い。安田・原（1982, 330頁）はこれらを総称してDK反応と呼び、DK反応の原因を4種類に分類している。①本来の意味でのDK（質問の意味が分からない，そのトピックについての知識がない）、②決められない（選択肢についてどちらともいえない，非日常的なことで想像がつかない）、③答えたくない（回答を拒否する）、④答えられないはずの人を調査員が誤って含めた、の4つである。これらのうち①や③は欠損データと言えるが、②や④は欠損データとは言えないであろう。

2）MARのより厳密な条件には、この未知パラメータξと欠損メカニズムのパラメータが独立である必要がある（Schafer 1997, pp.11）。

3）多重帰属推定は、ある欠損メカニズムをアプリオリに想定し欠損データの本来の値をシミュレーション的に推定する作業を複数回行う。その結果得られた推定値で欠損のない完全な素データを作成し、そのデータで必要な分析を行う。素データを作成するので、さまざまな標準的な分析手法を用いることができる。また、推定を複数回行うので、推定値の有効性の増加、推定の妥当性の確保（つまり欠損データによる「精度の問題」の正確な測定）、複数の欠損メカニズムに対する推定値の感度が分析可能、といった利点を持つ（Rubin 1987, pp.16）。最後の利点からも分かるように、多重帰属推定の根本原理は「無視できない」欠損メカニズムに基づく分析を含んだ非常に広い範囲をカバーできる可能性を持っているが、近年実際に運用されている多重帰属推定はMARに欠損メカニズムを限定されている。

なお、multiple imputationという名称はこの方法の内容を適切に表現していない、という批判がよく見かけられる。Wainer（1986, pp.4脚注）は「"parallel" or "repeated" imputation」という用語の方が適切

だと述べているし、岩崎（1993, pp.67）は「複数補填法といったほうがよいかもしれない」と述べている。しかし、multiple imputationの用語はすでに通用しているので、本稿でもこの用語に従う。「多重帰属推定」の訳語は筆者が当てたものである。岩崎（1993）は「多重代入法」という訳語を当てている。

4）欠損データの研究はある時期まで正規分布を仮定したものに限られていた（Hocking & Oxspring 1971）ので、カテゴリ欠損データを扱った研究は比較的珍しい。それでも、以前からいくつかの研究がある（Hartley 1958、Blumental 1968、Hocking & Oxspring 1971,1974）中で、Chen & Fienberg（1974）の方法が継承されているのは、一般化への可能性の高さを示すこれらの特徴があったためと考えられる。その1つの理由は本文でも述べたとおり、ログリニア・モデルとの接合が明確なことにあるが、もう1つの大きな理由として、パラメータの推定方法にEMアルゴリズム（EM algorithm）を使用していることがあると考えられる。現在、EMアルゴリズムの応用範囲は非常に広く、さまざまな研究分野で活用されている（MacLachlan & Krishnan 1997）。また、EMアルゴリズムの発展と欠損データの研究の歴史は密接に関係しており、欠損データの研究がEMアルゴリズムの結実を促したとも言われている（MacLachlan & Krishnan 1997, pp.38）。これらの意味で、パラメータの推定方法にEMアルゴリズムを用いていることは、Chen & Fienberg（1974）の方法が受け入れられるのに重要な役割を果たしたと考えられる。

5）ログリニア・モデルによる欠損データ分析は他にもいくつかの問題を持っている。適合するモデルが複数現れたときに最終的にどのモデルを選択するかの基準がはっきりしないこと（Fay 1986, pp.363、Baker & Laird 1988, pp.69）、「精度の問題」を無視していること（Baker & Laird 1988, pp.67-69、Park & Brown 1994）などは、比較的大きな問題である。他にも小さな問題としては、パラメータの推定値が境界解（パラメータ領域の端）に行き着いたとき、適切な最尤推定値を導き出せないこと（Baker & Laird 1988、Park & Brown 1994）、本来のモデル構成とは関係ないが欠損の発生には関わっていそうな変数を分析に加えるのが難しいことなどがある。

6）まず、Chen & Fienberg（1976）は自身の方法を多変数に拡張し、ログリニア・モデルとの親和性を明らかにした。Fuchs（1982）はその仕事を引き継ぎ、一部の変数のみが欠損を持つというよく遭遇するデータを扱う方法について一般化を施した。また、EMアルゴリズムの使用がここで初めて明示されている。

「無視できない」欠損データにまでこの方法を拡張したのはFay（1986）である。Fayは欠損データを扱う分析のほとんどが「無視できる」欠損データを仮定することで進められていることを批判し、欠損メカニズムをより的確に表現できる可能性がある「無視できない」欠損データを考慮する必要性を主張している。Fayはそれぞれが欠損を持つ2変数において「無視できない」欠損データを扱っていたが、Baker & Laird（1988）は欠損を持つ変数を1つに絞りながらも全体で扱う変数を多変量に一般化している。この研究で一般化の過程は完了したと言える。Baker & Lairdは推定値の標準誤差やモデルの選択基準について豊富な議論を行っており、その問題提起はその後の発展のために役立っている。

7）SSM調査で父親の学歴が欠損しているケースは、1955年調査では4.7%だったものが、65年には9.9%、75年には8.9%、85年では14.8%、95年では18.3%と上昇傾向を示している。ただし、55～75年データのサンプルは男性のみであるのに対して、85、95年データのサンプルには女性も含まれている。女性回答者の方が男性回答者よりもやや欠損が多く、欠損率上昇の一部はその影響である。ちなみに男性のみの場合、85年は13.7%、95年は16.6%である。

【引用文献】

Baker, Stuart G. and Laird, Nan M. 1988. "Regression Analysis for Categorical Variables With Outcome Subject to Nonignorable Nonresponse." in *Journal of the American Statistical Association* 83(401):pp.62-69.

Blumenthal, Saul. 1968. "Multinominal Sampling With Partially Categorized Data." in *Journal of the American Statistical Association* 63(322):pp.542-551.

Chen, Tar and Fienberg, Stephen E. 1974. "Two-Dimensional Contingency Tables With Both Completely and Partially Cross-Classified Data." in *Biometrics* 30(4):pp.629-642.

Chen, Tar and Fienberg, Stephen E. 1976. "The Analysis of Contingency Tables With Incompletely Classified Data." in *Biometrics* 32(1):pp.133-144.

Chen, T. Timothy. 1979. "Log-Linear Models for Categoricalized Data With Misclassification and Double Sampling." in *Journal of the American Statistical Association* 74(366):pp.481-488.

Tenenbein, Aaron. 1972. "A Double Sampling Scheme for Estimating from Misclassified Multinominal Data with Applications to Sampling Inspection." in *Technometrics* 14(1):pp.187-202.

Dempster, A. P., Laird, N. M., and Rubin, D. B. 1977. "Maximum Likelihood From Incomplete Data Via the EM Algorithm(With Discussion)." in *Journal of the Royal Statistical Society, Ser.B* 39(1):pp.1-38.

Fay, Robert E. 1986. "Causal Models for Patterns of Nonresponse." in *Journal of the American Statistical Association* 81(394):pp.354-365.

Fuchs, Camil. 1982. "Maximum Likelihood Estimation and Model Selection in Contingency Tables With Missing Data." in *Journal of the American Statistical Association* 77(378):pp.270-278.

Hartley, H. O. 1958. "Maximum Likelihood Estimation From Incomplete Data." in *Biometrics* 14(2):pp.174-194.

Hocking, R. R. and Oxspring, H. H. 1974. "The Analysis of Partially Categorized Contingency Data." in *Biometrics* 30(3):pp.469-483.

Hocking, R. R. and Oxspring, H. H. 1971. "Maximum Likelihood Estimation With Incomplete Multinominal Data." in *Journal of the American Statistical Association* 66(333):pp.65-70.

岩崎学. 1993.「多重代入法による欠落値の処理」『マーケティング・リサーチャー』 No.68:61-70.

Little, Roderick J. A. and Rubin, Donald B. 1987. *Statistical Analysis with Missing Data.* John Wiley & Sons, Inc.

McLachlan, Geoffrey J. and Krishnan, Thriyambakam. 1997. *The EM Algorithm and Extensions.* John Wiley & Sons, Inc.

Nordheim, E. V. 1984. "Inference From Nonrandomly Missing Categorical Data: an example from a genetic study on Turner's Syndrome." in *Journal of the American Statistical Association* 79(388):pp.772-780.

Park, Taesung and Brown, Morton B. 1994. "Models for Categorical Data With Nonignorable Nonresponse." in *Journal of the American Statistical Association* 89(425):pp.44-52.

Press, S. J. 1968. "Estimating From Misclassified Data." in *Journal of the American Statistical Association* 63(321):pp.123-133.

Rubin, Donald B. 1976. "Inference and missing data." in *Biometrika* 63(3):pp.581-592.

Rubin, Donald B. 1987. *Multiple Imputation for Nonresponse in Surveys.* John Wiley & Sons, Inc.

Schafer, J. L. 1997. *Analysis of Imcomplete Multivariate Data.* Chapman & Hall.

Wainer, Howard. 1986. "Introduction and Overview." pp.1-5. in *Drawing Inferences From Self-Selected Samples.* edited by Wainer, Howard. Springer-Verlag.

安田三郎. 1971.『社会移動の研究』 東京大学出版会.

安田三郎・原純輔. 1982.『社会調査ハンドブック〔第3版〕』 有斐閣双書.

保田時男. 2000.「クロス集計表における欠損データの分析: 学歴移動表を例として」『理論と方法』 近日中（Vol.15, No.1以降）に掲載予定.

The New Information Obtained by Missing Data Analyses —A Case of the Intergenerational Mobility in Education from 1995 SSM Survey—

Tokio YASUDA

Social scientists have been nervous about missing data because it causes a bias and/or low precision on the estimation from survey data. However, we could derive rich information from appropriate analyses of missing data. This paper discusses the significance of such a procedure, taking an example of intergenerational mobility in education. The log-linear model was used to compare the wide variety of missing mechanism including "nonignorable" one. The educational mobility table from fathers to children was analyzed; it had many nonresponses in father's category. Our method selected two models for missing mechanism, which had enough goodness of fit statistically. One suggests that younger generations receiving education under the postwar system have higher possibility of missing: the other suggests that sons of fathers having higher education have higher possibility of missing. These results add new information about the cause of nonresponses in father's education. Especially, the latter model means that missing group consists of movers downwardly. The information would be important for close analysis of intergenerational mobility in education.