

長期追跡パネル調査における継続・脱落の 要因分析

Analysis of the Sample Attrition in the Long Term Panel Survey

田磨 裕祐 TANABIKI, Yusuke (大阪大学大学院人間科学研究科博士後期課程)

パネル調査で発生するサンプル脱落は、得られたデータや分析結果の信頼性を左右する重大な問題である。本稿では、長期追跡パネル調査のサンプル脱落が、どのような個人属性や社会階層的要因によって説明できるのかを、社会調査の実務上の関心に基づいて明らかにする。その結果、居住の定着性や職業階層の影響を確認した。

キーワード 長期追跡パネル調査, サンプル脱落, 要因分析

1 パネル調査におけるサンプル脱落の問題

パネル調査とは、同一の調査対象者に対して繰り返して調査を行う、縦断的調査法 (longitudinal survey) の1つである。長期間にわたるパネル調査では、時間経過によるサンプルの脱落が不可避免的に発生する。これをサンプル脱落 (sample attrition) の問題とよぶ。

パネル調査データを用いた分析では、潜在的なバイアスを適正に評価し修正するために、サンプル脱落の問題が重要視される。パネルデータに限らず、横断的調査や繰り返し調査のデータを使用する場合でも、計画されたサンプルからの系統的な欠損がある場合、分析結果の一般化には慎重さが求められる。とくにパネル調査の場合は、サンプル脱落の機会が繰り返し生じるため、その要因について把握しておくことの重要性は高い。また、調査の実務的な関心からいえば、どのような対象者が脱落しやすいのかという知見の蓄積は、

調査資源や労力をどのように配分すればよいのかを知ることにつながり、有意義である。

本稿では、パネル調査におけるサンプル脱落が、諸個人のどのような属性や社会階層的要因によって規定されているのかについて、調査実務的な問題関心に基づいて検討する。したがって、サンプル脱落に起因するバイアスについての統計学的な検討は、本稿では行わない。

2 「職業とパーソナリティ調査」の概要

分析の対象として、1979年と2006年の2回にわたって実施された「職業とパーソナリティ調査 (以下WP調査と略記)」をとりあげる。¹⁾

WP調査の母集団は、1979年時点で関東7都県 (東京都, 神奈川県, 埼玉県, 栃木県, 群馬県, 茨城県, 千葉県) に在住していた有職男性である。2段階の層化無作為抽出法が用いられ、調査地点数は84地点、設計標本数は840であった。女性サンプルについては、男

性サンプルの配偶者に対して調査が行われた。年齢構成は、男性が26歳から65歳（平均43.4、標準偏差10.7）、女性が26歳から71歳（平均44.7、標準偏差10.2）となっている。第2波調査は、第1波から27年を経て実施された。年齢構成は、男性で53歳から88歳（平均66.8、標準偏差8.3）、女性で52歳から93歳（平均66.5、標準偏差8.4）となっている。

第2波調査は以下のように進められた。まず、第1波調査の有効回収サンプル（以下、ベースサンプル）に対する所在確認調査が行われた。これは、当時の住所宛に簡単なアンケートを郵送し、現在も同じ住所に居住しているのかどうかや、訪問による調査の協力を仰げるのかなどといった、対象者の現在の状況を確認するためのものである。つぎに、所在が確認できたサンプル（以下、捕捉サンプル）のうち、死亡や病気等といった理由で調査不能と判明したサンプルを除外し、第2波調査の対象とするサンプル（以下、対象サンプル）を選定した。つづいて実査が行われ、有効回収サンプル（以下、継続サンプル）を得た。

表1は、第2波調査のサンプル脱落の状況である。男女ともに、ベースサンプルの2割程度が、所在確認の失敗により脱落している。また捕捉サンプルの1割程度が、死亡・病気等の理由で調査不能となっている。第2波調査に限定した回収率は、男性が50.9%、女性が45.1%であった。ベースサンプルを基

準とした継続率は、男性が36.1%、女性が36.4%という結果となった。

WP調査は長期のインターバルを経た追跡調査であるため、現住所が確認できないことにより脱落したサンプルが多数存在する。したがって、所在確認時点での脱落要因が、重要な論点となる。また、所在確認時点での脱落と調査拒否による脱落の要因は異なりうるし、最終的な継続・脱落の確率は、それら個別の要因の複合的な結果であるとも考えられる。そこで本稿では、①所在確認時点での脱落、②実査時点での回答拒否、③最終的な継続と脱落、という3点について、個別に分析する。

3 サンプル脱落の要因分析

ここでは、サンプル脱落の発生要因をロジスティック回帰分析によって検討する。ロジスティック回帰分析は、従属変数が2値変数（本稿の場合は、サンプルの継続・脱落）である場合に、一般的に用いられる回帰分析である。サンプル脱落の発生確率（正確には、その対数オッズ）を、複数の要因を含めた回帰式によって説明することができるため、本稿の課題に適切な分析手法である。

説明変数には、第1波時点で得られた変数を利用する。具体的には、年齢、学歴、職業、従業上の地位、個人収入、世帯収入、居住地

表1 WP第2波調査のサンプル脱落の状況

男 性			女 性		
ベースサンプル	×	所在判明率	=	捕捉サンプル	
629		80.4%		506	
捕捉サンプル	×	健在率	=	対象サンプル	
506		88.1%		446	
対象サンプル	×	回収率	=	継続サンプル	
446		50.9%		227	
継続サンプル	÷	ベースサンプル	=	継続率	
227		629		36.1%	
ベースサンプル	×	所在判明率	=	捕捉サンプル	
418		85.4%		357	
捕捉サンプル	×	健在率	=	対象サンプル	
357		94.4%		337	
対象サンプル	×	回収率	=	継続サンプル	
337		45.1%		152	
継続サンプル	÷	ベースサンプル	=	継続率	
152		418		36.4%	

特性、配偶者の有無²（ダミー変数）、持ち家の有無（ダミー変数）を使用する。学歴については、旧制の学歴をもつサンプルも存在するため、カテゴリ変数ではなく連続的な教育年数を用いる。居住地特性は、東京・千葉・神奈川を「1」とし、その他の4県を「0」としたダミー変数を便宜的に用いており、居住地の都市化の度合いを代替的に表している変数であるとする。

先行研究から導かれる仮説は2つある（坂本，2006；宮内・McKenzie・木村，2006；Fitzgerald, Gottschalk and Moffitt, 1998）。まず、年齢的には若く、階層的に高い地位にあるサンプルが継続しやすいという仮説である。これらの要因は、先行研究において発見的に論じられてきたものであるが、なぜサンプル脱落に影響するのかについての統一的な説明はない。本稿では、年齢の効果は肉体的・精神的健康の度合いに由来し、階層的地位につい

ては、調査に対する理解や、調査協力につながる心理的な余裕を左右しているという解釈をしておく。第2に挙げられるのは、居住の定着性がサンプルの継続を促すという仮説である。結婚して世帯を構えることや、持ち家を所有することが、継続率を高めるのではないかという予測が成り立つ。

表2は、男性のサンプル脱落の要因について、2項ロジスティック回帰分析をした結果である。まず「継続／脱落」モデルは、最終的なサンプル継続・脱落の要因について分析したものである。次の「所在確認時点」モデルは、ベースサンプルの現在の所在確認に成功・失敗した要因についての説明である。最後の「実査時点」モデルでは、所在が把握できたサンプルにおいて、実査が成功・失敗した要因についての説明を試みている。

まず、所在確認時点で有意な正の効果をもつのは、「有配偶」と「持ち家」である。居

表2 サンプル脱落の2項ロジスティック回帰分析（男性）

	継続／脱落 ¹⁾		所在確認時点 ²⁾		実査時点 ³⁾	
	B	exp (B)	B	exp (B)	B	exp (B)
年齢	-0.07**	0.93	-0.01	0.99	-0.07**	0.93
教育年数	0.01	1.01	-0.03	0.97	0.00	1.00
専門・管理 (ref)						
事務・販売	-0.32	0.73	-0.08	0.92	-0.52+	0.59
技能・労務	-0.32	0.72	0.06	1.06	-0.59+	0.55
農業	-0.28	0.76	1.75	5.78	-0.33	0.72
経営者・役員 (ref)						
一般従業者	0.07	1.07	-0.46	0.63	0.20	1.22
単独・自営業主	-0.15	0.86	-0.46	0.63	-0.11	0.90
家族従業者	0.29	1.33	0.61	1.84	0.62	1.86
個人収入	0.00	1.00	0.00	1.00	0.00	1.00
世帯収入	0.00	1.00	0.00	1.00	0.00	1.00
都市部居住ダミー	-0.28	0.76	-0.37	0.69	0.01	1.01
有配偶ダミー	0.98**	2.66	0.91**	2.48	0.54	1.72
持ち家ダミー	0.72**	2.06	1.76**	5.80	-0.15	0.86
定数	1.44	4.21	1.32	3.75	2.99**	19.97
-2対数尤度		731.56		510.70		543.10
Cox & Snell R2 乗		0.11		0.14		0.13
N		609		609		434

** : p<.01 * : p<.05 + : p<.10

注) 各モデルの従属変数は以下のとおりである。1) 第2波調査を継続=1, 脱落=0,

2) 所在確認に成功=1, 失敗=0, 3) 対象サンプルのうちで、回収に成功=1, 回答拒否=0。

住の定着性を表すこれらの変数は、実査時点では有意な効果をもたず、所在確認時点と最終的な回収結果に影響を与えている。実査時点では、年齢と（「専門・管理」を基準カテゴリとした）「事務・販売」「技能・労務」が負の有意傾向を示している。高齢層やこれらの職業カテゴリに該当するサンプルは、所在確認が達成できたとしても、実査の段階で拒否する確率が高い。しかしながら所在確認時点では、これらの変数の効果は認められない。トータルでみたサンプル脱落に対しては、年齢の負の効果と、有配偶ダミーおよび持ち家ダミーの正の効果が認められる。いずれのモデルにおいても、疑似決定係数はそれほど高くはない。³

表3は、女性サンプルの「継続／脱落」モデルと「実査時点」モデルの結果である。女性サンプルについては、調査設計の段階で対象者を有職者に限定しておらず、無職が相当

数存在している。そこで、無職の効果を検討するために、「職業」と「従業上の地位」についての変数を、別々にモデルに投入した。また、欠損値が多くなるという理由から、個人収入は説明変数から除外している。第1波調査時点で、すべての女性サンプルに配偶者が存在することから、有配偶ダミーも分析に含めない。

「継続／脱落」「実査時点」という双方のモデルにおいて、一貫して有意傾向を示しているのは、年齢の負の効果と、職業・従業上の地位変数の正の効果である。職業変数では、基準カテゴリを無職に設定しているため、無職の脱落確率が高いと解釈できる。持ち家ダミーの正の効果は、「継続／脱落」モデルでのみ有意である。男性の場合と同じく、実査時点の脱落に対しては、有意な効果をもっていない。モデルの疑似決定係数はいずれも低く、女性においても、これらの変数の影響は

表3 サンプル脱落の2項ロジスティック回帰分析（女性）

	継続／脱落 ¹⁾				実査時点 ²⁾			
	B	exp (B)	B	exp (B)	B	exp (B)	B	exp (B)
年 齢	-0.07**	0.93	0.01**	0.94	-0.06**	0.94	-0.05**	0.95
教育年数	-0.02	0.98	0.06	0.98	-0.01	0.99	-0.03	0.97
無職 (ref)								
専門・管理	0.51	1.67			0.59	1.80		
事務・販売	0.51 ⁺	1.67			0.49 ⁺	1.64		
技能・労務	0.51 ⁺	1.67			0.74*	2.09		
農 業	1.33**	3.78			1.47**	4.36		
無職 (ref)								
一般従業者			0.27*	1.92			0.73*	2.08
単独・自営業主			0.55 ⁺	2.69			1.36*	3.90
家族従業者			0.30	1.51			0.38	1.47
世帯収入	0.00	1.00	0.00	1.00	0.00	1.00	0.00	1.00
都市部居住ダミー	-0.22	0.80	0.24	0.82	-0.24	0.79	-0.24	0.78
持ち家ダミー	0.93**	2.52	0.28**	2.64	0.42	1.51	0.50	1.66
定 数	1.54	4.65	1.08	3.38	1.66	5.26	1.48	4.39
-2 対数尤度		485.64		466.45		419.54		405.93
Cox & Snell R ² 乗		0.11		0.10		0.09		0.08
N		404		385		326		313

** : p<.01 * : p<.05 + : p<.10

注) 各モデルの従属変数は以下のとおりである。1) 第2波調査を継続=1, 脱落=0,

2) 対象サンプルのうちで、回収に成功=1, 回答拒否=0。

大きくはない。

4 長期追跡パネル調査におけるサンプル脱落

要因分析で得られた知見をまとめよう。持ち家や配偶者の有無といった、居住の定着性に関する変数が効果をもつのは、所在確認時点におけるサンプル捕捉の成否に対してである。一方で、職業的地位で代表される階層的属性の影響は、実査時点での回収の成否を左右している。これらの総合的な結果として、サンプル脱落は発生している。

このように調査手順の各段階によって、サンプル脱落の要因は異なっている。先行研究でも述べられているように、一般的に高い階層的地位にあるサンプルの継続率は高いが、それは捕捉が成功した後の実査時点においてのことである。捕捉の段階では、居住の定着性が低いサンプルの脱落にこそ留意する必要がある。すなわち、実査時点の回答拒否が多いパネル調査では、階層的属性の分布の偏りが発生する可能性が高まり、対象者の捕捉作業が不十分なパネル調査では、居住状況に関係する変数において偏りが生じやすくなるのである。以上の諸点は、われわれのような長期追跡パネル調査に固有のものではなく、調査回数の増加によって長期化したパネル調査に対しても適用できる知見であると考えられる。

本稿では、パネル調査で発生するサンプル脱落についての要因分析を行い、社会調査法の実務的な観点からの知見を示した。次になされるべき作業は、サンプル脱落によるデータのバイアスがどの程度であるのかという評価や、その修正といった、数理統計学的な検討である。これらの点を今後の課題とする。

注

- ・1 WP 調査を用いたこれまでの研究については、Naoi and Schooler (1985) や、吉川編 (2007) を参照されたい。なお WP 第 2 波調査は、文部科学省科学研究費の助成 (課題番号 19330106, 研究代表者: 吉川徹) による「職業と家族とパーソナリティについての長期追跡調査の日米比較分析」の一環として実施された。
- ・2 配偶者の有無の効果は、先行研究でさまざまに検討されている。「結婚予定ダミー」や「新婚ダミー」といった変数は、脱落を促す効果を示す (坂本, 2006)。しかし「有配偶ダミー」は、脱落を抑制する (宮内・McKenzie・木村, 2006)。これは家庭を構えることで、居住の定着性が低くなるからであると考えられる。本稿では、後者の効果を検討するために「有配偶ダミー」を使用している。
- ・3 継続・脱落の発生確率に対しての、第 1 波調査時に得られた変数の説明力が低いという結果によって、データに系統的なバイアスが存在していないという判断ができるのだろうか。脱落以降の観察不可能な変数にも依存する非ランダム脱落 (北村, 2005) の存在が想定できる場合、その判断は保留せざるをえなくなる。現実的には、とくに長期にまたがるパネル調査では、一定の非ランダム脱落が存在していると考えerるほうが自然であろう。このような脱落の存在は、長期追跡パネル調査の限界であるといわざるをえない。この点についての統計学的な検討は、別稿にゆずりたい。

文献

- Fitzgerald, J., P. Gottschalk and R. Moffitt, 1998, "An Analysis of Sample Attrition in Panel Data: The Michigan Panel Study of Income Dynamics", *The Journal of Human Resources*, 33 (2): 251-99.
- 吉川徹編, 2007, 『階層化する社会意識——職業とパーソナリティの計量社会学』勁草書房。
- 北村行伸, 2005, 『パネルデータ分析』岩波書店。
- 宮内環・C. R. McKenzie・木村正一, 2006, 「パネルデータ継続と回答行動の分析」樋口美雄・慶應義塾大学経商連携 21 世紀 COE 編『日本の家計行動のダイナミズム II —— 税制改革と家計の対応』慶應義塾大学出版会, 9-52。
- Naoi, A., and C. Schooler, 1985, "Occupational Conditions and Psychological Functioning in Japan", *American Journal of Sociology*, 90 (4): 729-52.
- 坂本和靖, 2006, 「サンプル脱落に関する分析——『消費生活に関するパネル調査』を用いた脱落の規定要因と推計バイアスの検証」『日本労働研究雑誌』48 (6): 55-70。