特集論文

「消費生活に関するパネル調査」の変数管理について

坂口 尚文
（公益財団法人 家計経済研究所 次席研究員）

1. 調査の継続により発生している問題

「消費生活に関するパネル調査」（以下、JPSC）は、家計経済研究所が日本初の全国規模のパネル調査として1993年から開始し、以降、毎年実施してきた調査である。本稿執筆時の最新調査である2011年の調査は、19回目の調査にあたる。調査が20回近く継続したことで、データから得られる情報量はクロスセクションデータ20回分以上の価値をもっている。一方で調査の長期継続に伴う制度疲労が随所にあらわれているのもまた事実である。

パネル調査の長期継続に関わる問題としては、対象者が脱落していく、サンプルの摩耗（attrition）がよく知られている。摩耗はサンプルサイズの減少にとどまらず[1)]、サンプルの偏りの問題を引き起こす要因ともなりうる。データの代表性の観点から、脱落をなるべく減らし、どのような属性の対象者が脱落しているかを把握することがJPSCにおいても常々求められてきた[2)]。

対象者数のように調査を継続していく過程で減少するものがある一方、調査とともに増加・増大の一途をたどるものがある。それはデータの規模である。規模もまた、違った角度から問題となっている。規模の大きいデータのハンドリングは、利用と管理の両面に負担をかける。パネルデータは過去のデータを包含しつつ、成長を続けるデータである。当年のデータを前年のデータと比べると、物理的容量と調査からの情報量がともに増加する。物理的容量の増大については、調査開始時は分析、保存の両面においてネックとなる要素ではあったのだが、ハードウエアの進歩により、今ではさほど問題とはならなくなっている。これまで19回の調査を繰り返しても、観測数は人・年ベースで33,000弱である。個人のパソコン環境を考えても、同規模のデータを扱うことはさほど難しくないことがおわかりいただけるだろう。

情報量の増加の方は大きく分けて、対象者の異時点の情報が追加されたことにより獲得されたもの、そして調査票に新たに加えた質問項目から獲得されたものの2つがある。データ管理の面でより問題になっているのは、質問項目の増加である。本稿ではこの質問項目の増加がJPSCのデータ管理、提供の面においてどのような形で問題になっているのか、そしてどのような対策を考えているかについて簡単に紹介する。あくまでJPSCの個別問題であるが、すべてのパネル調査が抱えうる問題に敷衍できる内容もあるかと思う。

2. 質問項目の増加で何が問題になっているのか

調査に含まれる質問項目が多いことは、調査がそれだけ豊かな情報を有していることになる。分析の用途幅も広がるため、データの有用性を示す重要な指標の一つとも言える。ただ、質問項目の数が多すぎることは、データのハンドリング面で深刻な問題を引き起こす。一言で言えば、「何がどこにあるのか分からない」状態である。情報量はエントロピーを指標として表されるが、質問

項目の毎年の追加により情報量が増加していく様相は、物理のエントロピー増大則と同様に、データが無秩序へと向かっている感さえある。過去のデータを包含するパネルデータの特性上、情報量（エントロピー）は増加こそすれ、決して減ることはないのである。

話を具体的な方向に移そう。調査の質問項目はデータ上では、変数という名目で格納されている。JPSCのデータに含まれる変数の数は、第1回調査から第19回調査までの延べで4,706に上る。パネル調査は繰り返し調査であるため、全く同一の質問が含まれる。同一内容の質問には、各調査回で同一の変数を割り当てているため、JPSCには異なる変数が4,706個存在していることになる。だが、データの利用者が4,706個の変数をすべて使って、分析のモデルを構築することはない[3]。4,700の1%を考えても47である。これまでのJPSCを用いた論文をみる限り、一回の分析で40から50程度の変数を用いるケースは、変数の使用が多い部類に入る。つまり、ほとんどの利用者にとっては、JPSCに収録されている変数の99%以上は自分の分析には必要のないものと言える。「藁山の中から針を探す」ということわざがある。自身が必要とする変数を効率的に探し出す方法がないならば、JPSCの利用者はそのような状況に置かれていることになる。

変数の数が膨大となり、全体像の把握が困難なことはデータを管理する面でも問題である。データの作成・管理者であるわれわれは、毎年、実査および分析を行っている。そのため使用頻度が高い変数や長期継続して尋ねている質問については、その変数名や質問文の内容についておおよそ把握できるようになっている。ただ、使用頻度が低いからといって、調査上、重要でない変数とは言えないことが問題である。例えば、消費税の引き上げの影響については、以前の引き上げの調査内容との比較が必要とされる。また、意識項目などの近年あらためて注目を浴びるようになった項目は、10年以上の時を経て調査に再録されている。

JPSCではなぜ変数の数がここまで増大しているのか、次節ではその背景について整理してみる。

3. パネル調査は変数が増加する性質を持っている

調査の質問項目と変数の関係は、一対一の関係の場合が主だが、複数回答のように一対多の関係の場合もある。いくつかの質問項目をまとめて事後的に変数を作成すれば、多対一の関係にもなりうる。いずれにせよ、質問項目には対応する変数が存在するため、データに含まれる変数の多さは調査票に含まれる質問項目の多さと考えて、ほぼ差し支えないだろう。

もちろん、調査から得られたデータに含まれる変数が多い[4]ことは、JPSCをはじめとするパネル調査だけの特徴ではない。クロスセクションの調査でもボリュームが大きい調査は数多く存在する。そのような調査ではデータに含まれる変数の数も多い。ただ、全国規模の社会調査に限定すれば、パネル調査というだけで、調査に質問項目が多く含まれることの十分条件足りうる。変数の数が膨張する性質がパネル調査には内在しているからである。

なぜパネル調査では変数の数が膨大となるのか。その原因の一つとして、調査の継続によって変数が付加されていくことを簡単に触れたが、すでに初回調査の時点で膨大な変数を含んでいることも見落としてはならない。この初回調査の変数をベースに新規の質問を追加、修正していくからである。初回時、継続の別に個々の要素について以下に詳しく述べよう。

まず、初回調査時点で膨大な変数を含む点についてである。ここでも大きく分けて2つの要因があげられる。1つは、パネルデータはデータが自己完結している必要があるためである。パネルデータを用いた分析では、同一の対象者による複数時点の計測値に意味がある。パネル調査のデータは集計しない形、すなわち個票（マイクロデータ）ベースで公開し、かつ利用することが前提となる。個人レベルの情報は他の調査のデータから外挿することは難しい。特にJPSCのように対象をランダムに抽出し、かつ匿名化しているデータでは外挿は不可能である。そのため、個人の属性や行動

に関する情報は同一調査でくまなく収集する必要がある。

調査の初回時点で変数の数が膨大になるもう1つの要因は、調査実施に伴う組織的な要因である。パネル調査は金銭的、また時間的にもコストがかかるプロジェクトである。一般論として、パネル調査は同程度の規模のサンプルを持つクロスセクション調査よりも、一回あたりの実施費用は高くなる。特に大きな割合を占めているのが、対象者に支払う謝金の存在である。上述したように調査票に多くの設問を盛り込むため、その回答量への対価も高く提示する必要がある。また、対象者に調査に継続して参加するインセンティブを持たせるため、通常、一時点限りの調査よりも高額の謝礼を払う傾向にある。この他､エディティング等､データの事後的な品質管理に割く費用、転居等に伴う対象者の追跡にかかる費用も無視できない要素である。ただ、費用面で事実上、最も大きな負担となるのは、一回あたりの調査費用自体よりも、継続的に調査を実現するために安定した財源の確保と運営組織を確立、維持していくことである。

パネル調査の実施にあたっては費用が莫大なものとなるため、個別の研究者がパネル調査を単独で実施することは効率的でない。ゆえに、一つの調査にさまざまな分野の研究者が相乗りして運営する体制を、国内国外問わず、ほとんどすべての調査プロジェクトがとっている。研究関心が異なる者が多く参加することは、どうしても質問項目の増加を招く。JPSCでも、調査の運営方針は20名近くが参加する研究会の合意で行っている。このことが質問項目の増加の背景にあることは事実である。JPSCでは60数ページの調査票を対象者に配布し、各調査年のデータには1,300から1,500程度の変数が含まれている。ただ、JPSCの調査票は紙の調査票である。回答者の（心理的）負担を考慮して、一回60ページを超えないようにする暗黙のルールがあり、そのことがシーリングの役目として変数増加の抑制に一定の効果を果たしている。それでも19回調査時点で変数の総数が4,706であるから、一回あたりの調査票の3ないし4倍程度の量の変数増加を許容したことになる。

次に、調査の継続に伴い変数が増加していく要素について述べよう。パネル調査は同一個体の異時点の値を収集することが目的である。そのため、同一の内容の質問を同じ形式で尋ね続けることが大前提となる。一方で、JPSCはパネル調査という以前に社会の実態を捉える調査でもある。調査期間中も社会情勢は刻一刻と変化し続けるため、社会の変化に応じ、実態に即した質問を適宜、追加および修正する必要がある。例えば、ある政策の立案を何年も前から予知することは不可能であるし、その政策の影響を把握するニーズがなくなれば該当項目を調査からはずし、新たな項目を追加した方が調査全体の有用性は高まる。

また長いスパンで捉えると、調査開始当初に設定した質問や選択肢の内容が社会の実態にそぐわないものになることがある。例えば、労働者の雇用形態などは、この20年で多様化が進展した事例であり、選択肢の内容を調査の途中で変更している。この場合は、選択肢のコードが違う以上、同一の質問内容であっても、変更前後で別の変数として扱わざるを得ない。調査期間の長さは、社会環境だけでなく対象者自身の変化も伴う。JPSCは特定年齢を対象にしたコーホート追跡の調査である。調査開始時に20歳代後半であった対象者も､20年が経つと40歳代後半になっている。就業、結婚、出産、子育て、いずれの面においても40歳代には20歳代の時点とは違う問題にも直面している。コーホート追跡調査の場合には、年齢に則した質問を適宜追加する必要がある。このようにパネル調査では質問内容の新陳代謝を繰り返し、結果として変数の数は累積していく。

JPSCでは、第1回から第19回すべての調査回に含まれている変数の数は318である。データに含まれる変数の総数は4,706であるから、9割以上の変数が調査の途中から導入されている。全国規模のパネル調査では、程度の差こそあれ、変数の増加が避けられないと割り切る必要がある。そうであるならば多量の変数の識別、そして変数と質問項目の対応づけ（マッピング）をどのように行うかが鍵となってくる。変数識別の第一歩は変数の名前づけである。

4. 変数管理の実際

(1) 変数名について

JPSCの変数名の命名規則はきわめて単純である。原則、Q123のような通し番号で管理している。番号は変数に対応する質問項目のデータ上での出現順（格納順）に合わせ、昇順に発行している。番号の値自体に意味はない。2回目以降の調査では、新規の質問項目に関しては、前年の番号からの続きの番号を振っていくが、同一の質問項目には引き続き同じ変数名を与える。同一の質問項目であっても選択肢など質問の方式を変えた場合には、新しい変数名を発行する。なお、回答がマルチアンサーの場合は、個々の選択肢に該当/非該当の2値の値を振る。そのため、選択肢の数だけ変数を割り当てている。この場合、同一設問内での連番の数字は共通にして、接尾にアルファベットを順に充てて区別している。選択肢1→Q123A、2→Q123B、……といった具合である。これら昇順・連番の規則から大きく逸脱する変数は、対象者を識別する"ID"という変数と何回調査のデータであるかを示す"PANEL"という変数だけである。

昇順に管理と書いたが、事情を有り体に言えば、調査開始時点から変数命名のグランドデザインがあって、明確な目的意識から昇順に発行したわけではない。むしろ、最も安易な方式を選んでいたと言える。ただ、変数の数が膨れ上がった現状から振り返ると、変数名に意味を持たせないシンプルな命名規則であったからこそ、結果的にデータの管理・提供システムが破綻せずに今日まで継続できたといえる。

JPSCで昇順の命名規則を採用していたメリットは、以下のような点であったといえる。

1. 名前の一意性を担保でき変数名の衝突が確実に避けられる。
2. 変数名をそのつど考案する必要がないため、リリースまでの時間を短縮できる。
3. 調査票の構造に依存していないので、調査票、質問内容の変更による影響を受けない。

1、2については、クロスセクションデータにも当てはまる話で、パネルデータ固有の話ではない。変数名に意味を持たせる場合は、データ作成者の間で命名規則について意思疎通が必要である。パネル調査の場合は変数名の発行時点だけでなく、将来の担当者との意思疎通も考える点が特筆すべき要素かもしれない。直接の後任だけでなく、調査継続期間、担当する者を含めてである。将来担当者が変数名および規則を解読する有象無象の学習費用は無視できない。命名の範囲を絞り、重要な変数だけ意味のある名前を持たせるという考え方もある。ただ、何が重要で何が重要でないかの判断は、時間の経過とともに移ろいやすいものである。初期時点で重要と思われていた項目が重視されなくなることもあるし、後々、重要と考えられるようになる項目もある。一度つけた変数名は、調査期間中は変更することが難しい。なお2については、JPSCでは海外ユーザーにもデータを提供している。誰にとっても意味のある名前にするならば英語の表記を考える必要があり、そのコストもあながち無視できない。

3の変更への頑健性は、パネル調査ならではの事項といえる。前節で述べたように、パネル調査も社会情勢の変化などに応じて、設問そのものや選択肢の内容が変更されていく。異なる質問文、選択肢から得られたデータは別の変数とみなすため、変数名も違うものとなる。表意名として完結していた名前を、さらにどう改定するかは難しい問題である。接尾に改定を示す文字、たとえばnewを追加するにも、では3度目の改定があった場合はどうするのかといった問題はつきまとう。命名法については、その変数がどのカテゴリーに属する質問項目かだけでも名前に盛り込もうとする方式もある。変数名の接頭か接尾にキーワードを埋め込む。あるいは特定の番号、例えば500番台をそのカテゴリー用に確保して、一種の名前空間を作りだす方式である。しかしJPSCでは大項目のレベルでも質問の廃止は往々にあった。その際、当該項目に属するすべての項目が廃止されるならばよいが、一部の項目は存続し他の大項目に統合されることもある。残存項目は最新調査とは違うカテゴリー名を付与されたままとなり、経緯

を知らない者への混乱を招く。

変数名の発行時点に将来の調査内容、質問項目の変化を予測することは不可能である。一時点で変数名を最適化すると変化に対応できない。かといって、変数の命名規則に従って、調査票設計の自由度を制限することは本末転倒である。変数名の一時点の最適化は、時点時点で結果が確定しているクロスセクションデータでは望ましい対応であっても、パネルデータでは必ずしも適切な対応とは言えないのである。

プログラミングの世界では変数にできるだけ意味のある名前をつけることが推奨されている。これは、プログラマーが変数の参照されるスコープの大きさを、自ら調整できる余地があるからである。JPSCのデータに含まれる変数の集合は5,000近くの要素数をもつ。この要素数は所与である。質問の内容には類似した項目、あるいは付問の項目も多い。意味のある名前をつける場合、それらを識別するためにも名前が長くなることは必須である。変数が保持する内容（セマンティクス）を変数名という限られた数文字に押し込むことは合理的でない。分析者にとっても、データの変数名はテンポラリーなものとして割り切り、自分のプログラム内で自分にとって意味のある名前に変更した方が効率的である。

このように、JPSCでは変数名にデータを識別するID以上の機能を持たせていない。質問項目と変数の対応はデータの外で構築する。その方針自体は上記に述べた理由から正当化できると思う。ただ、データの複雑性に見合う、変数と質問項目の対応機能をわれわれは用意していたかと言われれば、答えは否である。調査の開始から20年近くが経過し、変数の数が膨大化したため、そのことの問題が表面化している。

(2) 変数と質問項目をどう対応させているか

現在、JPSCでは質問項目と変数の対応を表形式で整理して提供している（図表-1）。1列目に変数を記し、該当する質問項目はその変数と同じ行に記入している。そもそも、このJPSCの対応表は固定長データのレイアウト表である。JPSCのデータは、固定長レコードのデータで調査会社から納品されている。固定長レコードは桁数で変数の位置（アドレス）を指定する。そのアドレス部分を削除し､代わりに変数を付与したにすぎない[5]。

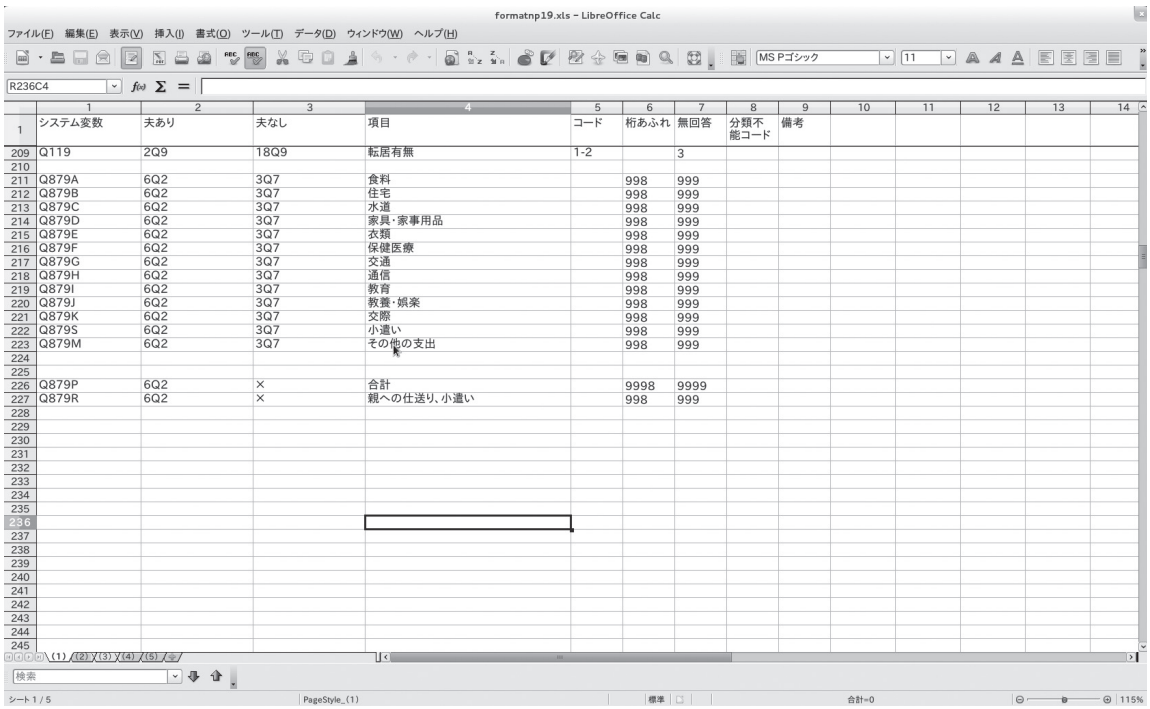
対応が図表-1のような表形式であることによる、利用者側のメリットはほとんどないのだが、強いてあげれば、表計算ソフトはJPSCの利用者のほとんどすべてが所有していること、それゆえ、ソフトの扱いや表形式の構造に慣れていることである。つまり、誰でもファイルが開けて、記載されている内容も理解、類推はできる。実際、多くの調査でも表形式の対応表が提供されているため、利用者にとっても既視感がある[6]。

対して表形式のデメリットは、件数がある程度以上に達すると可読性が下がること、この一点に尽きる。どの程度の件数とは一概に言えないが、全体を一瞥できない範囲、つまりスクロールが必要な範囲になると可読性は確実に下がり始める。可読性の低さは、閲覧者が提示された情報を処理し切れていない状況に他ならない。

ここまで、変数と質問項目の対応しか述べてこなかったが、質問項目に付随する情報は変数名だけではない。分析にあたって最も必要とされるのは、変数の値が示す内容である。年齢や年収については値から類推できるかもしれないが、幸福感の質問では1という値を見ただけで、その内容は類推できない[7]。変数の値の情報については、無回答に割り振るコードの情報や上限値（桁あふれ）に割り振るトップコードについての情報も含まれる。このようなさまざまな変数の値の情報を対応表に加えるのは困難を伴う。技術的には対応する列を加えるだけなので問題はないのだが、列数が増えることは対応表のさらなる可読性の低下を招く。表形式は情報の追加に強い構造ではない。

また、表形式は枠（セル）に簡単な文字列しか情報を書き込めない問題も有している。こちらも技術的な問題ではなく、可読性の問題である。質問項目には、年齢、性別といった項目名だけで事足りるものもあれば、内容を把握するためには調査票の質問文を要するものもある。質問文が長い場合は、枠に記入することが難しい。さらに質問

図表-1　表形式による対応



	1	2	3	4	5	6	7	8	9	10	11	12	13	14
1	システム変数	夫あり	夫なし	項目	コード	桁あふれ	無回答	分類不能コード	備考					
209	Q119	2Q9	18Q9	転居有無	1-2		3							
210														
211	Q879A	6Q2	3Q7	食料		998	999							
212	Q879B	6Q2	3Q7	住宅		998	999							
213	Q879C	6Q2	3Q7	水道		998	999							
214	Q879D	6Q2	3Q7	家具・家事用品		998	999							
215	Q879E	6Q2	3Q7	衣類		998	999							
216	Q879F	6Q2	3Q7	保健医療		998	999							
217	Q879G	6Q2	3Q7	交通		998	999							
218	Q879H	6Q2	3Q7	通信		998	999							
219	Q879I	6Q2	3Q7	教育		998	999							
220	Q879J	6Q2	3Q7	教養・娯楽		998	999							
221	Q879K	6Q2	3Q7	交際		998	999							
222	Q879S	6Q2	3Q7	小遣い		998	999							
223	Q879M	6Q2	3Q7	その他の支出		998	999							
224														
225														
226	Q879P	6Q2	×	合計		9998	9999							
227	Q879R	6Q2	×	親への仕送り, 小遣い		998	999							
228														
229														
230														
231														
232														
233														
234														
235														
236														
237														
238														
239														
240														
241														
242														
243														
244														
245														

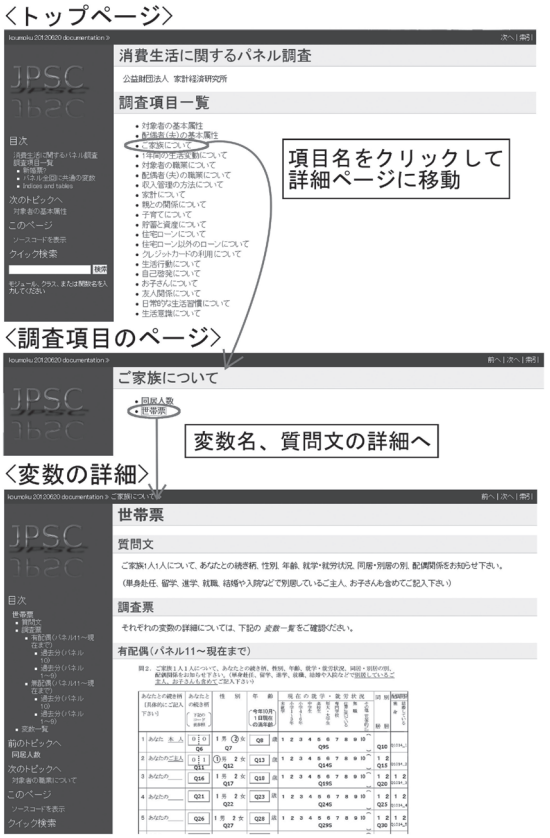
文に場合分けが含まれるものなどは、その構造を枠に押し込むことはできない。この問題の解決案として、調査票に直接変数名を埋め込む方式を検討したことはある。確かに変数の値と質問文・選択肢のリンクという問題はクリアする。ただ、調査票は各年単位で分割されているので、変数の変遷は把握することはできない。最大の欠点は、利用者が調査票を読むことを前提にしていることである。JPSCの調査票は一冊あたり60～70ページにわたる。さらに有配偶票と無配偶票の別がある。データを分析したいだけの利用者に20年近くの調査票を精査させることは現実的ではない。

前述したように、パネル調査では同一の質問項目であっても、選択肢の変更等により変数名が変遷していることがある。時間の要素が含まれるパネルデータを2次元の表に押し込むことは理にかなっていない。

(3) 改善に向けて

インターネットの発達に伴い、注目を集めてきた言葉として「情報アーキテクチャ」というものがある。内容を一言で言い表すのは難しい言葉だが、「発信者がどうすれば情報をうまく伝えられるか」、そして「受信者が情報をどうすればうまく探し出せるか」を表現する技術や形式である。特にWebページの構成のあり方を議論する場面で言及されることが多い。Morville and Rosenfeld（2007）は、情報アーキテクチャがなぜ重要であるかについていくつかの理由をあげている。その中に「情報を見つけるためのコスト」と「情報を見つけないことによるコスト」を抑えられる点がある。表形式で変数を羅列する方式はこれら2つのコストが高いといえる。Morville and Rosenfeldがあげた理由で、さらに興味深いのは、「教育の価値」である。商品を探しにWebページを訪れた顧客に関連商品や新商品を教育するといった文脈で用いられているが、JPSCにも参考になる点がある。検索した変数に関連する項目を一覧で提示すれば、利用者のモデルの改良に寄与できる可能性がある。また選択肢の変更など変数

図表-2　木構造による対応



の改定があった場合は、以前の変数への言及があれば利用者の利便性は高まる。

現在、JPSCで着手しているのは、変数の集合に木構造を入れて表現することである。木構造とは要素を親と子の関係で結んでいく方式である[8)]。木構造というと難しく聞こえるかもしれないが、図表-2のように個々の設問へリンクをたどっていく階層構造だと考えればイメージがわくだろう。なぜ変数の集合に木構造を取り入れて整理するかといえば、調査票に木構造を見出すことができるからである。JPSCの調査票は20程度の大項目があり、その中に関連する個々の設問が配置されている。例えば、対象者の職業→現在の職業→労働時間といった具合にナビゲートしていくことができる[9)]。木の先端（葉）には質問項目と変数の対応を記述するファイルを用意する（図表-3）。1変数に1ファイル割り当てれば、スペースの確保ができ可読性の問題は解決できる。変数の詳細な情報も多く記載することも可能である。

木構造で質問項目の変遷も容易に表現できる。項目を統合するダミーの枠をつくり、その配下に図表-3のようなファイル（葉）を新旧個々、水平に並べていけばよい。JPSCでは同一の質問項目であっても有配偶、無配偶で質問内容が微妙に異なるケースもある。こちらも表形式では対応が難しいケースであったが、木構造であれば「有配偶」と「無配偶」の具体的項目を水平に並べて追加するだけで同様の対応が可能である。ただ、木構造を用いる一番のメリットはwebとの親和性かもしれない。利用者がいつでもアクセス可能なインタラクティブな調査票を提供できる。

最後にデータ利用者の利便性という原点に立ち返れば、利用者が最終的に必要としているのは変数名ではない。自分が使用したい独自のデータセットである。変数名はその媒介役にすぎない。利用者に自分が必要な変数を選んでもらい、自身が必要とするだけのデータを提供するフレームワークを考えていく必要がある。フレームワークの要件を考えると、利用者が何度もデータをリクエストできる双方向性は必須である。利用者は分析を進めながら必要な変数を取捨選択していくのが通常である。さらに、開発の容易さ、オペレーティング・システムに非依存という要素を考慮すれば、Webアプリケーションによる公開が優良解である。実際、アメリカの代表的なパネル調査であるPanel Study of Income DynamicsやNational Longitudinal SurveysのWebサイトでは、すでにこの機能が実現されている。一方で、データをインターネット上にあげることは、セキュリティの問題に直面する。統計データのプライバシー保護という観点からは重要な問題である。

これら改善に向けた対応は現在、作業・構想段階である。実際の作業に伴う具体的な課題は、別の機会に改めて論じられればと思う。

図表-3　変数の詳細

5. おわりに

日本におけるパネル調査の先駆けとして、JPSCはパネルデータを収集することこそが第一の使命であった。現在もそして今後も、そのことが重要であることに変わりはない。ただ、調査開始から20年近くたった今、蓄積したデータを資産としてどう管理、活用していくかについても視線を向けざるを得ない状況になっている。

データが家計経済研究所の私的なコレクションであれば、調査の詳細は実施者のみがノウハウとして会得しておけばよく、利用者に対してもデータを使いたいならそのノウハウを時間をかけて覚えなさいと言える。だがJPSCは公共性のあるデータであることを標榜している。つまり多くの分析によって、意義ある知見が得られるよう努め、その知見を広く社会へ還元できることを目指している。公共データの提供者は私的な収集家ではなく、たとえれば、誰もが使いやすい公共の図書館的な役割を果たさなければならないだろう。図書の整理・保存・提供は図書館機能の根幹をなす。それらの機能が充実しているからこそ、利用者は他事にとらわれることなく自身の研究に専念できるのである。OPACの導入により図書館における検索の利便性・効率性は著しく高まった。カードの目録が図書館から姿を消したように、JPSCにおいても表形式の変数対応表は同様の運命をたどるべき存在かもしれない。

最後にこのような形で図書館のアナロジーを出したのは、本稿で述べたトピックは図書館情報学の世界では、中心テーマの一つとして長く議論されてきたトピックだからである。膨大な量の書籍が厳然として存在しており、それらを管理する必要に迫られて発展してきた領域ともいえる。もちろん大量の情報を扱うことは図書館に限ったことではなく、他のさまざまな分野においても確認できる。それらを包括する情報工学での近年の急速な発展や議論を見れば、膨大な情報を扱うことの重要性の認識と各個別分野に付随する困難を解決していく姿勢がみてとれる。

一方、日本におけるパネルデータの管理の現状は、情報関連分野の成果や研究者層の厚さに比べて、見劣りするのは否めない。日本ではパネル調査の歴史が浅く、データの蓄積がいまだ発展途上であるため、データの管理が明確な問題点として認識されていないのは致し方ないところである。データ利用者の絶対数が、図書館などに比べて圧倒的に少ないことも、データの管理やユーザビリティーへの対応が後回しになった要因でもあろう。

形はどうあれ、そう遠くない将来、日本においてもパネルデータの収集にコンピューターを導入する時期はくるだろう。データ収集時のコンピューターの使用は、回答者の回答に対して同時点のみならず異時点の論理的不整合を即時に修正できる。そのため、データの質の面で大きな変革をもたらすことが期待できる（保田 2012）。さらに、調査へのコンピュータの導入は、過去も含めた対象者の属性や回答に応じて質問の変更が可能になる。データから得られる情報量が増大する反面、質問の項目総数やデータ構造の複雑性もさらに増すことが予想される。そのようなデータを受け入れる素地はわれわれにできているのか。これからの調査のあり方を考えつつも、これまでの調査結果についても情報の管理・提示法を再考する必要がある。

補論　変数と変数名

本文中では「変数」と「変数名」という言葉を区別せずに用いた。プログラミングの世界では、両者はやや異なる概念だが、JPSCのデータを用いた論文や利用手引きや申請書類などでは、両者の違いが明確に意識されることは少なく、議論も複雑になるため本文中では意図的に区別しなかった。

プログラミングにおける変数とは、データを記憶しておくための（一時的な）領域を指している。プログラミング言語によって若干の差異はあるが、データを格納しておく入れ物とか場所と思ってよいだろう。変数名はこの入れ物につけられたラベルである。入れ物に入っているデータは変数の値と呼ばれる。

JPSCの文脈で変数をこのような記憶領域として捉えることはまずない。JPSCのデータはデータであって、コンピューターに対する処理を記述したプログラムではないため当然とはいえる。JPSCのデータは作成された段階で、すでに変数に名前がつけられ、値が代入されて永続化した状態である。そのため、明示的に記憶領域が意識されず、代わりに存在が意識できる変数名と変数の値（観測値）が変数という言葉で語られるのかもしれない。さらに、話を少しややこしくしているのは、JPSCの文脈で変数といった場合、年齢や年収などの質問項目やその概念とほぼ同義語として用いられることである。こちらは、統計モデルでの独立変数、従属変数からの援用であろう。質問項目と、そのデータ上の実体である観測値、観測値を表す名前が「変数」という言葉で、混同、同一視されている。

パネルデータはしばしば3次元の構造を持ったデータであると言及される。その際、次元を構成するものの一つが「変数名」の空間である。JPSCでは、対象者の識別ID、調査回数、変数名の三要素を指定すれば、値が一意に定まる。例えば、ID "2000"番の人の第"10"回目の調査時の年齢（"変数名：Q8"）は、"30"歳といった具合である。次元の意味を、純朴に個々の点を特定するために必要な要素数と考えれば、3つ組の値（2000,10,Q8）と30の間に対応が成り立つ3次元のデータと捉えることができる。

ただ、この座標と値の対応に学問的な価値はほとんど見出せない。JPSCではランダムサンプリングにより対象者を抽出し、対象者の匿名性を担保している。対象者のIDが持つ情報は個体の識別だけである。ID2000番の対象者の年齢が分かったとしても、そのこと自体がもつインプリケーションはほぼ皆無である。そもそも、プライバシー保護の観点から、特定個人の情報に焦点を置いた分析をデータの利用規約で禁止している。

JPSCを使った分析の多くは線型モデルを主体としている。つまり、質問項目間の足し算やスカラー倍に学問的見地から何らかの意味をもたせ

た、「質問項目」あるいはその「概念」の空間である。実際の分析は、各質問項目を対象者、および調査回数という座標で測った「観測値」を使って展開する[10]。この場合、線型空間の次元という意味では、モデルに使用した質問項目の数が次元となる。

注

1) JPSCの毎回の脱落率は5%程度であるが、それでも20回近くの調査を重ねると累積で半数近くが脱落している。
2) 例えば、JPSCの脱落等に関する研究として重川（1997）、村上（2003）、坂本（2006）がある。
3) あるとしたら、パネルデータのメタ分析だろうが、JPSCではいまだ行われていない。
4) 変数の数が多い少ないというのは主観的な表現ではある。例えば、代表的な表計算ソフトであるExcelでは、Excel2003以前のバージョンで列数を256に制限していた。また、いくつかのデータベース管理システムでは、表のカラム数制限をデフォルトで1,000あたりにおいているものがある。少なくとも、含まれる変数の数が1,000を超えるような調査は変数の数が多い範疇に入るだろう。
5) データのレイアウト表であるため、各年各年、その調査の対応表を提供している。19回調査の現在、19回調査分＋3回の新規対象追加時のデータ分の計22枚のファイルがある。ファイル数が多く、うまく統合がなされていないことも検索効率の悪さに影響を与えている。
6) 実際、国内のパネル調査の多くでも、表形式で提供されている。
7) ちなみにJPSCで幸福感は5段階評価であり、1は「とても幸せ」を示す。
8) パソコンのディレクトリー構造を想像することが、一番イメージがつかみやすいかもしれない。
9) 木構造での親子の関係は、親と子の包含関係としても捉えることができる。包含による集合の分類である。JPSCで変数の数が多いといっても高々5,000程度の話である。個体認識は困難でも分類は可能な数字である。インターネット上に散らばる情報は分類さえも困難な状況に達している。
10) パネル調査の場合は、時間軸を示す調査回数も準主役として、質問項目の調査回ごとの観測値を基底として扱うケースは多い。単純化したイメージで言うなら、時間が座標扱いのケースは異なる調査回のデータを縦に結合した場合で、準主役のケースは横につなげた場合である。

文献

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』551: 55-70.

重川純子，1997,「「消費生活に関するパネル調査」における欠票の特性」『季刊家計経済研究』33: 76-83.

村上あかね，2003,「なぜ脱落したのか」財団法人家計経済研究所編『家計・仕事・暮らしと女性の現在――消費生活に関するパネル調査 第10年度』国立印刷局, 115-122.

保田時男，2012,「パネルデータの収集と管理をめぐる方法論的課題」『理論と方法』27(1): 85-98.

Morville, Peter and Louis Rosenfeld, 2007, *Information Architecture for the World Wide Web*, Sebastopol, CA: O' Reilly Media.

さかぐち・なおふみ　公益財団法人 家計経済研究所 次席研究員。主な論文に「母親の教育期待とその推移」（『季刊家計経済研究』88，2010）。労働経済学専攻。（sakaguchi@kakeiken.or.jp）