第2講　パネルデータの調査方法と構造

2.1　はじめに

パネルデータを利用した研究やその統計手法の研究は増加の一途をたどっているが、パネルデータの調査方法に関する研究は、それらに比べれば、それほど進んでいるとは言えない。そんな中でも、欧米の 300 人を超える専門家が 1986 年 11 月にワシントンＤＣに集まり、パネルデータの調査方法に関する国際会議がアメリカ統計学会を主要なスポンサーとして開かれた。その主要な論文は Kasprzyk *et al* (eds)(1989) に収録されている。そこに収められている課題は次の通りである。(1) パネルデータ調査のデザイン問題、(2) パネルデータ収集と調査票デザイン問題、(3) 統計上の問題と推計、(4) データベース管理、(5) 統計誤差の問題、(6) パネルデータ調査の時点間調整の問題（前回の回答に規定される等）、(7) 非回答調整、(8) パネルデータによる因果関係分析、(9) パネルデータ・モデルの意義。このうち (1) ― (7) までがパネルデータ統計調査の問題に関わるものである。とりわけ、調査方法上の問題として、調査対象の選択範囲 (coverage) と非回答 (nonresponse)、脱落（attrition）が重要であることが指摘されている。アメリカの代表的なパネルデータ調査である Panel Study of Income Dynamics (PSID) では 1968 年に 4802 家族で調査を始めたが、1989 年には累積で 50 %を超える家族が非回答となっていることからもわかるように、パネルデータ調査が長期化するに従って、調査の最初に選択された標本から多くの参加者が脱落していき、標本と母集団の関係性が歪んでくることが最大の問題となっている。

我が国のパネルデータ調査に関する解説は林知己夫（編）(2002)『社会調査ハンドブック』に収録されている松田年弘「パネル調査」(pp.262-267) があるが、体系的な研究は進んでいない。もちろん、これには我が国においてパネルデータ蓄積の歴史が浅く、利用実績も限られており、調査としてどのような問題が出てきているかさえ十分には理解されていないという事情がある。

本章ではこれまでに明らかになってきているパネルデータ調査に関わる主要な問題点およびその解決方法について論じ、後半ではパネルデータ調査から得られた情報を基に、具体的にパネルデータセットをどのように作ればいいのかを解説したい。

2.2　調査方法

パネルデータ調査も基本的には統計調査法に基づいて設計され、実施されている[1]。具体的な内容は調査によって違ってくるが、一般的な流れとしては次のようになっている。

[1] 統計調査法の詳細については、例えば、松田・伴・美添 (2000、第 1-3 章)、豊田 (1998)、鈴木・高橋 (1998) などを参照されたい。

（i）調査目標の設定と調査票の作成

調査の目的と具体的な調査内容が設定され、それに応じて誤解なく答えられるように、後で解釈に困るようなことのないように、適切な質問項目を作る。その際、平均的な回答時間を考慮して質問の数や配列を決め、また個人属性に関する質問にも十分配慮することが必要である。調査票には調査対象者の属性に関する質問（**フェイス項目**と呼ぶ）が含まれる[2]。調査の内容に関わる項目は、回答内容を予め選択肢として用意した**プリコード項目**と質問に関する回答を自由に記入させる**自由記述項目**に大別できる。プリコード項目は回答しやすい反面、回答の選択肢が限定されており、予想外の答えには対応できないという問題がある。具体的には、多肢選択法（一つだけ選択）、無制限複数選択法（複数選択可）、制限複数選択法（選択数を制限）、完全順位法（全てに順位を付ける）、一部順位法（一部選択して順位付けする）などがある。自由記述項目は数値記入法と文字記入法がある。統計分析では数値記入法によるデータが最も用いられるが、質的情報としてプリコード項目の選択肢を用いることもある。質問の設定の仕方には気を付けるべき点もあるし、回答拒否を誘うような質問は回避し、事実と評価を区別することも重要である。

(ii) 調査仕様の決定

調査の対象となる母集団を決め、抽出単位が個人か世帯か法人かを決める。調査対象を母集団全てとする場合には**全数調査**（**悉皆調査**）、一部を取り出す場合を**標本調査**という。全数調査には多大な費用と労力がかかるので、一般に標本調査が行われる。また調査期間や費用、調査対象となる標本数も決める必要がある。パネルデータを収集するためにはそれなりのノウハウが必要であり、実際には経験を積んだ各省庁統計担当部局あるいは民間調査会社が仕様を決めて、調査の予算に応じて標本数が決まっているようである。

（iii）標本の抽出

標本抽出の方法は**無作為抽出法**(random sampling) と**有為抽出法**[3](purposive selection) の 2 つがあり、一般には無作為抽出法が用いられることが多い。母集団の中から標本として等確率で抽出されるような方法を**無作為抽出法**とよぶ[4]。この無作為抽出の方法はさらに分類することができるが、ここでは代表的な**層別抽出法**(stratified sampling) と **2 段抽出法**（two-stage sampling）について見ておきたい。層別抽出法は属性の構成比率の予備知識を利用して母

[2]属性を表わす代表的なフェイス項目としては、性別、年齢、学歴、職業、年収、労働経験年数、婚姻形態、住居形態、住所、同居家族、兄弟姉妹の有無、職歴、父母の状況、子供の数などがある。

[3]有意抽出法としては知人、同僚などに調査協力を依存する機縁法・紹介法、モニターに応募してきた人を調査する応募法、典型的な標本を選ぶ典型法、街頭で調査するインターセプト法、専門知識を要約していくデルファイ法、母集団の構成比に等しくなるように標本を集める割当法などがある。

[4]後節で論じるようにパネルデータでは標本が時間を経るに従って徐々に脱落していく問題がある。これが特定の属性の標本に見られる現象であれば、抽出時点でこの特定の属性の標本の抽出率を高めておくといった処置が望ましいだろう。しかし、どのような属性の標本が脱落しやすいかを事前に把握することは難しいし、日本の脱落の事例を見ても、調査開始後の状況の変化によって脱落するということが多いので、標本抽出の段階で脱落問題に対処するのは実際には難しい。

集団を層別し、各層に対して、**住民基本台帳**などの**台帳**（フレーム）から乱数を用いて標本を抽出する。層別化しない**単純無作為抽出法**より精度が高いとされている。2 段抽出法は母集団を地域によって 1 次抽出単位（都道府県、市町村等）に分け、まず 1 次抽出単位を抽出した後で、その単位から標本を抽出する。この 2 つの抽出方法を組み合わせた**層別 2 段抽出法**では母集団をいくつかの層に分け、層ごとに 2 段抽出を行うものであり、全国規模の主要な調査がこの方法を採用している。母集団から標本を抽出するためには、母集団全ての対象が含まれている**台帳**が必要になるが、世帯であれば、国勢調査や住民基本台帳に基づくことが多い。企業であれば、事業所統計調査、工業統計調査、商業統計調査などの企業センサスや国税庁資料に基づく営利法人名簿、帝国データバンクのデータベース等に基づいて抽出されることが多い。

（iv）調査票による調査の実施

調査票に答えてもらうためには一般に次のような方法がとられることが多い。 **面接調査** この調査は調査員が直接調査対象者のところまで出向き、調査への協力を依頼する。この方法では回答がその場で得られるため確実に情報が取れるし、調査員が記入するので間違いも少ない。しかし、限られた時間で質問するために質問項目は限定され、記憶に頼るような質問に対しては誤差が入ってくる可能性が高い。 **郵送調査** この調査では調査票を郵送で送りつけ一定期間後に郵送によって回収するものであるが、対象者の時間の都合に応じて答えられるし、その分時間も多少かかるような質問もできる。しかし、対象者の自主性に依存しているために回収率は一般に低い。 **電話調査** この方法は対象地域の電話帳から無作為に電話番号を抽出して対象者を選ぶものであり、調査費用は上の 2 つと比べるとかなり安くつく[5]。しかし、電話帳から無作為抽出したとしても、対象地域に住む住民が特定の属性や時間帯によって回答者が偏る場合には問題となる。 **留置調査** 調査員が調査対象者の住居を訪問し、調査の主旨を説明し、調査票を配布し一定期間内に記入しておくことを頼む。一定期間後、調査員が調査票を受け取りに再び訪問し、調査票に記入漏れがないかどうかを確かめた上で、調査票を回収する。この方法は面接法に近いが一定期間時間を置くので、かなり多項目の調査も可能になる。また、調査員が複数回訪問することで、信頼関係が生まれることも調査にとってはメリットである。パネルデータ調査のように継続的に同一対象者に調査を依頼する場合には、費用はかかるが、この留置法によることが望ましい。 **集合調査** これは一定の場所に対象者を集めて回答してもらう方式で、学校、会社、病院など人が集まってくる場所が対象になることが多い。これも標本の代表性という意味では問題があるが、一度に回答が回収できるということで費用節約的な調査方法である。

（v）調査結果の編集・集計

回収された調査票には様々な誤差が入っており、データとして入力する前にそれらの誤差を出来る限り修正する。単純な記入ミスや回答方法の誤解な

[5]アメリカで電話調査が多用されるのはこの費用の問題が大きいと思われる。

どで適切な回答が類推できる場合には修正を施す。質問とは関係のない答え方をしているものなどは無回答扱いとする。これらの作業は個々の問題に当たってみなければ一般的な解決方法が在るわけではない。調査実施者は修正に恣意性が入らないように厳密な手続きを決めておく必要がある。有効な調査票を標本数で割り、**回収率**を計算しておくことも重要である。またパネルデータでは累積した脱落者数や**脱落率**も把握しておくべきである。

パネルデータは具体的には次の 3 種類の調査方法によって集められている。

（i）クロスセクション調査で調査対象が複数回の調査で重複しているケース

このタイプの調査はパネルデータを作成する目的で行われたわけではなく、一定の条件を満たす経済主体が必ず調査対象となるようにデザインされたものである。例えば、証券取引所に上場しているすべての企業は『有価証券報告書』を財務省に提出する必要があるが、『有価証券報告書』の企業財務データを同一企業について複数年つなぎ合わせれば企業のパネルデータを作ることができる。同様に経済産業省で調査している『企業活動基本調査』は資本金 2000 万円、従業者数 50 名以上の製造業を中心とするすべての企業を調査対象としている。この調査でも、従業者数が 50 名以下になるか、廃業するのでなければ必ず繰り返し調査対象になるので、事後的にパネルデータとして再構成することが可能になる[6]。

これらの調査では一定の条件を満たす主体がすべて調査の対象になるという意味では標本調査ではなく（ある種の全数調査・悉皆調査）、調査対象が途中で理由もなく脱落するという問題もほとんどない。しかし、各年の連続性は意識されておらず、回答者も年によって交代することも多いので、場合によっては、回答者の理解の違いや記入誤差によってデータが大きくぶれることもある。このタイプのデータを利用する場合には、データの非連続性が本当の変化なのか誤差なのかを注意深く吟味する必要がある。

（ii）クロスセクション調査で調査対象が一定期間継続して調査に参加し、一定の割合で調査対象が入れ替わるケース

この調査はパネルデータ調査と考えることもできるが、一般的にはクロスセクション調査として設計されている。具体的な例としては総務省の『家計調査』がある[7]。この調査では 6ヶ月間同一の家計が家計簿をつけ、毎月 6 分 1 のサンプルが入れ替わる。詳細な家計簿を 6ヶ月連続して付けることには、調査対象にかなりの負担を強いることになるが、海外の同様の家計調査ではインタビュー形式ではるかに短い期間（例えば 1 週間分）の消費について調査しているのに過ぎないことと比べると信頼のおける調査となっている。また毎月の調査の連続性という意味でも 6 分の 5 が前月と同じ家計であることから、標本の交替による不連続性は小さい。

この調査は全国の二人以上の一般世帯および単身の一般世帯を対象に、全国の市町村から調査市町村を抽出し、抽出された調査市町村から調査単位区を抽出した後に、調査単位区の中から調査世帯を抽出する層化 3 段抽出法に

[6] この調査を用いた実証研究は第 7 章を参照。
[7] この調査を用いた実証研究は第 8 章を参照。

よっている。調査世帯の交替は 1 調査単位区 6 世帯を単位として全国で毎月 6 分の 1 ずつ行っている。

（iii）パネルデータ調査で調査世帯の交替は行わない

これは当初より同一主体を継続的に調査し、統計を蓄積することを目的に設計されている。このタイプの調査としては財団法人家計経済研究所の『消費生活に関するパネル調査』がある[8]。この調査は 1993 年から毎年実施されており、1993 年時点で満 24 歳から 34 歳の 1500 人の女性をコーホート A として、1998 年時点で満 24 歳から 27 歳の 500 人の女性をコーホート B として追跡調査している。調査は全国の都道府県を 8 ブロックに分類し、そのブロックを都市規模によって 13 大都市、その他の市、町村に分類した層化 2 段無作為抽出を行っている。また調査票を留め置いて一定期間後に回収するという留置法を用いているのでインタビュー形式に見られる記憶違いなどの問題は少ない。しかし逆に時間が経つにつれてサンプルが脱落していく問題はある。この点については後述する。

2.3 調査上の問題点

既に何度も論じてきたように、パネルデータ調査は同じ対象を繰り返し調査するということでここの対象の時間を通じた変化を捉えることができ経済行動を分析する上で非常に有益ではあるが、そのことは同時に時間を通して回答者集団の母集団に対する代表性が次第に失われていくという問題を抱えていることも意味している。

この問題はいくつかの理由で生じる。第一にパネルデータ調査に対して慣れてくることによって回答に歪みが生じる可能性がある。第二に以前の調査の回答に縛られて正直な回答ができないということも考えられる。第三に次第に調査に参加することがわずらわしくなり回答拒否（脱落）するようになり、第四に転居などによって追跡が難しくなるということも考えられる。

パネルデータ調査において代表性が確保されているかどうかは 5 つのレベルで検討されるべきである[9]。

（i）標本設定時脱落による歪み

パネルデータ調査では母集団から無作為抽出した標本に対して、調査に先立ってモニターの受諾を確認する。この時点で拒否されるケースを標本設定時脱落という。この結果、脱落した標本が以後のパネルデータ調査にもたらす歪み（偏り）を測定することは難しい。というのは脱落した標本からは一度も調査を行っていないからである。しかし抽出過程で性別、年齢、地域などの住民代表ベースの情報が用いられていれば、それを用いて調査不能になったグループと調査回収できたグループを比較し、調査不能グループに対しても調査に対する回答を予測（推定）することが可能になる。この推定結果と

[8] この調査の実証研究は第 9 章を参照。

[9] 以下の議論は松田年弘「パネル調査」（林知己夫（編）（2002）『社会調査ハンドブック』に収録）を参照している。

実際のパネルデータ調査の結果を対照すれば、調査不能グループのもたらした歪み（偏り）が推計できる。このように、標本設定時に脱落したり、調査のかなり初期に脱落するグループに対しては母集団の同位置層から代替標本を無作為抽出して補填することが多いが、その新たに選んだグループが調査不能グループの歪みを補正していることを確かめることが必要になる。

（ii）継続時脱落による歪み

ある程度、調査を継続したのちに何らかの理由で脱落する標本もある。これがまったくランダムに発生しているのであれば大きな問題ではないが、脱落が一定の理由によるシステマティックなものであれば、それは問題を含んでいる。この歪み（偏り）を評価するためには、途中で脱落した調査継続標本が脱落する前までに回答していた数値を調査継続して脱落してない標本と比較することで、その歪みを評価することができる。

（iii）調査慣れがもたらす歪み

調査慣れや以前の調査の影響によって回答にどの程度歪みがもたらされているかは、新たに無作為抽出した標本と比較することで評価できる。この評価のために新たな標本を導入するということは費用もかかるし、実際の手間も大きい。この種の歪みが大きいとわかっている場合には、調査自体に慣れを生じさせないような工夫、過去の調査の影響を少なくする質問の仕方を考えるべきであろう。

（iv）回答者の同一性の確認

パネルデータ調査を訪問留置法によって行う場合、標本抽出された本人ではない他の家族が回答するケースも見られる。具体的な数値データであれば、矛盾に気づくことも多いが、意識調査に対する回答に別人の回答がパネルデータとして入ってくると、深刻な誤差を生じさせることになり、この問題に対しては回答者が本人であるかどうかの確認を調査票に入れることが重要である。

（v）回答誤差

パネルデータ調査で同一の質問を複数回に亘って行う場合、回答に誤差が見られることがある。もちろん本当に意見が変わる場合もあるだろうが、回答者が違っていたり、回答時点での心理状態が違うといったことも考えられる。パネルデータ調査における各回のマージナル分布が同じで、前回と今回のクロス表がほぼ対称という条件を満たすならば、各回答者には本来の態度があり、態度の強度に応じていくつかの集団に分類され、態度強度が異なると質問に対する回答選択の確率が異なる、各質問に対する回答誤差は独立である等を仮定して回答誤差を推計するモデルを構築することができる。

2.4 脱落サンプル問題

上述の調査上の問題点の中でも、標本が一定期間後に脱落していくケースは広範に見らるが、この問題が検討されることは、これまで統計実務家など一部の関係者に限定されていた。しかし、近年、パネルデータの利用が増えるに従い、またパネルデータの蓄積が進むに従い、脱落サンプルの問題は認知

されるようになってきた。実際、*The Journal of Human Resources* の 1998 年春号（vol.33, no.2）がパネルデータ調査の脱落サンプル問題を特集しているし、Fitzmaurice, Laird and Ware (2004) の教科書でも 1 章を割いて（第 14 章）この問題を論じているように、近年、計量経済学者や統計学者、様々な分野の実証研究者の間でこの問題に関心が集まっている。

2.4.1 脱落サンプルの実態

まず限定的ではあるが、脱落サンプルが実際のパネルデータ調査でどの程度起こっているのかを確認しておこう。パネルデータ調査の先進国であるアメリカでは代表的なパネルデータ調査である The Panel Study of Income Dynamics (PSID) に関して脱落サンプルの問題が詳細に検討されている（Fitzgwerald, Gottschalk and Moffitt (1998a, b) , Moffit, Fitzgerald and Gottschalk (1999)、Lillard and Panis (1998)、Ziliak and Kniesner (1998)）。彼らの研究によると、1968 年には 4802 家族が標本抽出され[10]、翌年には 88 %の家族が残り、12 %が脱落している。以後、1989 年に至るまで年率 2.5 − 3 %が脱落し、1989 年時点で 49 %の家族が残り、51 %が累積して脱落していった。脱落の理由は転居、死亡、家族全体の非回答などが挙げられている。脱落者の属性を分類すると、（1）社会扶助、（2）未婚者、（3）高齢者、（4）有色人種、（5）低学歴、（6）労働時間が短い、(7) 低賃金、（8）借家住まい、を満たす家族である可能性が高いことがわかった。これは、脱落者が一般的には社会的に低い地位にいる可能性が高いということを意味している。しかし、脱落者の中には高所得者も含まれていることも確認されている。

The National Longitudinal Survey of Youth (NLSY) の脱落サンプルについては MaCurdy, Mroz and Gritz (1998) が詳しく検討している。NLSY は 1957 年 1 月 1 日から 1964 年 12 月 31 日までに生まれたアメリカ国民から標本抽出したもので、（1）6111 の一般若者家計、(2)1480 のラテン系若者家計、（3）2172 の黒人若者家計、（4）経済的に困窮している 1643 の白人若者家計、（5）1280 の軍関係の若者が含まれている。調査は 1979 年から毎年実施されている。1991 年時点で、累積の脱落率は男性に関しては、全体で 11% 、白人 10.5% 、黒人 11.7% 、ラテン系 11.3% と極めて低い。女性に関してはさらに低く、全体で 8.0% 、白人 7.3% 、黒人 8.4% 、ラテン系 9.1% となっている。脱落しやすい属性としては、男性では (1) 無業、(2)20 代の勤労者（賃金・所得には無関係）、（3）10 代の勤労者で賃金所得の高い層、女性では (1) 無業、(2) 学歴はあまり脱落には関係がない、（3）10 代の勤労者で賃金の高い層、20 代ではその差は消滅する、（4）高卒、大卒の賃金所得の高い層、となっている。ここから共通して言えることは無業者に脱落が多いことと、10 − 20 代で脱落している人の所得はむしろ高いということである。これは、PSID の結果とは異なっている。

[10] 4802 家族の内、2930 家族（5 分 3）がミシガン大学の The Institute of Social Rearch 附属の Survey Reaerch Center (SRC) の台帳（フレーム）から選択され、残りの 1872 家族（5 分の 2）が The Bureau of the Census の Survey of Economic Opportunity (SEO) に含まれている低所得家計（SMSA に入っている）台帳から選ばれた。

Zabel (1989) は PSID と The Survey of Income and Program Participation (SIPP) の脱落率を比較している。PSID の初年度から次年度にかけての脱落率が 12% であったのに比べて 1984 年に始まった SIPP は次年度で 6% であり、7 年後の 1990 年でも 71.4% が残っていると報告されている。

Newmark and Kawaguchi (2001) は The Current Population Survey (CPS) が移動した人を追跡しておらず、脱落サンプル問題が発生しているという点を指摘している。CPS と同じ標本抽出のフレームを用いている SIPP が移動した人を追跡する努力をしていることに目を付け、CPS と SIPP を統計的にマッチングさせ、CPI に関して脱落サンプルバイアスの調整を行った。労働組合の賃金効果には脱落サンプルバイアスはほとんど見られないのに対して、男性の結婚プリミアム（高賃金）には統計的に有意なバイアスが見出された。しかし、その額は経済的に意味のある程大きなものではない。総合的に判断すると、CPS の持っている情報量は脱落サンプルバイアスを凌駕するものであって、パネルデータ調査の価値は決して下がるものではないと結論づけている。

Burkam and Lee (1998) はアメリカ教育省が集めている High School and Beyond Study (HS&B) というパネルデータを分析している。この調査は 1980 年に 12000 人の高校生（約 1000 の高校から無作為抽出）を対象に、卒業間際の成績とその後の職業生活への影響を調べる目的で行われたものである。国立教育統計センターの努力により 78% のサンプルが 4 回の調査全てに回答している。脱落率は非常に低い。

Van den Berg and Linderboom (1998) はオランダの Labour Supply Panel Survey を使っている。この調査はオランダの正規学生を除いた 15 － 61 歳の 4020 人の個人（2132 家計）を対象に労働に関する情報を集める目的で 1985 年から始められたもので、1990 年までに 4 回の調査が行われている。1992 年時点で、元のサンプルにいた個人の 34% にあたる 1384 人が継続して調査に参加しており、残りの 2636 人（64% ）が脱落したことになる。1986 年、1988 年、1990 年に追加サンプルを加えて、1992 年時点でサンプル数 4438 人を確保している。

我が国のパイオニア的パネルデータ調査である『消費生活に関するパネルデータ調査』（家計経済研究所）でも脱落サンプル問題が出ており、平成 15 年度版『家計・仕事・暮らしと女性の現在：消費生活に関するパネルデータ調査（第 10 年度）』の第 III 部で、この調査がどの程度問題となっているかを論じている。この調査は 1993 年より始まっており、2003 年で 10 回の調査が行われているが、その内訳は，1993 年時点で 24 － 34 歳 1500 人で始まったコーホート A と 1998 年度から 24 － 27 歳 500 人で始まったコーホート B に分かれる。その内、脱落サンプルはコーホート A で 471 人、コーホート B で 175 人、合計 656 人となっている。脱落比率に直すとコーホート A が 31.4% 、コーホート B が 35% となっている。この数字をアメリカの PSID の 10 年目の数字である 30.3% と比べるとほぼ同程度の脱落率であることがわかる。

村上（2003）は脱落理由を分析している。過半数を占める「詳細不明」を

別にすると、コーホート A では「多忙」(25.1%)、「転居先不明」(12.5%)、「(長期) 不在」(10.6%) となり、コーホート B では「多忙」(24.1%)、「転居先不明」「(長期) 不在」「結婚」がそれぞれ 12.0% となっている。さらにデータを詳しく見ると、調査の初期の段階では「家族の反対」が多く、時間がたつにつれて就職・転居・結婚などのライフイベントの発生によって脱落していく傾向がある。復活したサンプルと復活しなかったサンプルを比べると、「死亡」「転居先不明」などのように物理的に不可能な場合、調査そのものに不信感、負担感がある場合は復活していない。「結婚」「離婚・別居」「転居」「家族の反対」などの理由を挙げる人も復活していない。「結婚」「家族の反対」は結婚相手が調査に反対するケースが多いと言われている。これは質問が本人のみならず、結婚相手やその両親にまで及ぶためであると思われる[11]。「離婚・別居」「転居」を機に脱落するのはいろいろな意味で心機一転したいので、継続的な調査もやめてしまいたいということであろう。逆に、復活しうるのは「(長期) 不在」「入院中」「体調不良・病気」「多忙」「出産」「就職・転職」「家族の病気・不幸」など脱落理由が一時的なものである場合に限られる。

坂本 (2003) は脱落サンプルの統計的特性を考察している。方法としては「継続回答者」と「脱落者」を分け、前年の回答データの平均値を比較している。それによると、脱落しやすい属性は、(1) 年齢が若い、(2) 未婚者、(3) 低学歴、(4) 有業者、(5) 高所得者、(6) 子供の人数が少ない、(7) 結婚予定者、(8) 核家族、となっている。これらの理由は一部アメリカの脱落理由と重なるが、有業で高所得者ほど脱落しやすいというのは逆である。また結婚を機に調査から脱落するというのもこの調査の特徴となっている。

脱落サンプルの問題は先進国のパネルデータ調査だけではなく、開発途上国のパネルデータにも当てはまる。ここではいくつかの途上国のパネルデータの脱落サンプル問題の実態について把握しておきたい。

この問題に取り組んでいる研究としては、Thomas, Frankenberg and Smith (2001)、Alderman, Behrman, Kohler, Maluccio and Cotts Watkins (2000)、Maluccio (2000) などがある。Thomas, *et al* (2001) では The Indonesia Family Life Survey (IFLS) の脱落サンプルいついて論じている。この調査は生活に関する実態を様々な側面から調査しようという目的で継続的に行われている。第 1 回調査は 1993 年に行われ、13 地域で 7000 を超える世帯が調査された[12]。その後、1997 年に再調査が行われたが、アジア金融危機の影響もあり、1998 年にも一部のサンプルについて危機の影響を知るために再々調査が行われた。第 3 回目の大々的な再調査は 2000 年に行われた。第 2 回の再調査に答えたサンプルは第 1 回調査の 93.5% であり、脱落率はわずか 6.5% にすぎない。死亡者を除けば、脱落率はさらに下がり 5.6% になる。この数字は途

[11] 日本の場合、広い意味で結婚を理由に脱落しているサンプルが無視できないほどある。結婚前後の労働供給や出産、育児などの行動を分析することは、女性パネルデータ分析では焦眉となるトピックであり、それに該当するサンプルが脱落していくことは統計調査としても重要な問題である。村上 (2003) が指摘しているように、対象者およびその家族に調査の意義を理解してもらい信頼関係を築くこと、対象者が多忙な場合にも回答が得られるような工夫をすること、対象者に感謝の気持ちを伝えることなど、地道な努力が必要である。

[12] このサンプルの基となったフレームは全人口の 83 %をカバーしている。

上国のパネルデータ調査としては極めて低い[13]。この低脱落率の理由は対象者の移動後先を追跡・確認して再調査しているためであるとされている。

Alderman, Behrman, Kohler, Maluccio and Cotts Watkins (2000) が用いたボリビアの Bolivian Pre-School Program Evaluation Household Survey Data (El Proyecto Integral de Desarrollo Infantil: PIDI) は子供の栄養摂取や知的発育が親の労働を確保するプログラムによってどの程度改善されるかを、都市部の幼児のいる世帯を対象に調査したものである。この調査の 1995 年と 1998 年の 2 年間隔の調査での脱落率は 35 %であり、ケニアで行われた The Kenyan Ideational Change Survey (KDICP) は、避妊や AIDS に関する情報や行動が非公式のネットワークを通してどのように伝わるかを見るために、4 つの地方村落で行われたパネルデータ調査である。この調査の 1994/5 年と 1996/7 年の 2 年間隔の調査での脱落率は夫婦世帯で 41% 、男性世帯で 33% 、女性世帯で 28% であると報告されている。ナイジェリアのパネル家計調査では 5 年間隔で 50% の脱落率があった。南アフリカでは 1993 年に行われた南アフリカ最初の全国規模の家計調査 (the 1993 Project for Statistics on Living Standards and Development(PSLSD)) があり、2 段抽出法がとられていた。その中で最も人口密度の高い KwaZulu-Natal 県の PSLSD に参加した標本に対して 1998 年に再調査を行ったのが KwaZulu-Natal Income Dynamics Study (KIDS) である。1993 年に 1393 世帯あった標本のうち、5 年後の 1998 年には 1171 世帯が再調査に回答した。この脱落率は 16% と極めて低い。修学以前の幼児のいる世帯の脱落率は 22 %となっている。

これらの途上国のパネルデータ調査の脱落理由としては移動のために追跡が不可能になったというものが大きい[14]。他には、不在、回答拒否、病気、多忙、死亡、離婚・別居などの理由が挙げられているが、移動と比べるとはるかに小さい。

Maluccio（2000）は南アフリカの KwaZulu-Natal Income Dynamics Study (KIDS) を取り上げ、脱落に関するケース・スタディーを行っている。この調査の脱落率が低い原因として、移転した世帯を適切に追跡し、再調査できたことが大きいとされている。これによって脱落率が 25% は押さえられたと言われている（標本の 4.5% ）。

このように、パネルデータからの脱落という同一の行動を見ても、それぞれの国、それぞれの調査によって脱落理由もパターンも異なっていることが明らかになった。しかし、脱落率の低い調査はいずれも調査機関が調査対象に対して積極的にアプローチし、移動しても追跡調査するなど、かなりの努

[13]他に、脱落率の低いパネルデータ調査としては、中国の健康栄養調査 (The China Health and Nutrition Survey) は 1989 年に 3795 家計を調査し、2 年後に再調査した時には 95 %の同一家計から回答を得たという事例がある。Alderman *et al* (2000) によると、インドの 1970/71 年の地方の世帯調査は 11 年後の 1981/82 年に再調査した時でも脱落率が 33% とこれも長い間隔があるにしてはそれほど高くないことが示されている。マレーシアの調査でも 12 年の間隔があるにもかかわらず脱落率は 25% であると報告されている。アジアの国で脱落率が比較的低いのには様々な理由があると思われるが、統計調査に対する信頼・協力という側面が大きいと言われている。

[14]例えば、ケニアの KICS では、男性の 47.8% 、女性の 37.9% が移動によって追跡が不可能になっている。

力をしていることが見て取れる。これはパネルデータ調査を継続的に行う上で重要な点であり、この眼に見えない努力がパネルデータの利用価値を高めていることを認識しておく必要があるだろう。

2.4.2 脱落サンプル・バイアスの識別

次に、脱落サンプルが多く存在し、時間の経過とともにその数が増えていくとすると、そのようなパネルデータを用いた推計のバイアスを確定することが必要となってくる。

まず、統計分析に入る前に、脱落のメカニズムを整理しておこう。一般に脱落は次の 3 つのケースに分類できる。(1) **完全ランダム脱落**(Missing Completely At Random: MCAR)　この場合には脱落によるバイアスは無視できる。(2) **ランダム脱落**(Missing At Random：MAR or Selection on Observables)　脱落は脱落以前までの観察可能なデータによって推測できる。(3) **非ランダム脱落**(Missing At No Random: MANR or Selection on Unobservables)　脱落が脱落以後の観察不可能なデータにも依存しており、観察可能データのみによって脱落を推測することは難しい。

統計的に問題がないのは (1) のケースであり、(2) は観察可能なデータを用いて何とか対処できる。(3) は対処が極めて難しい。

また脱落サンプルを含んだパネルデータをどのように扱うかという観点からも分類することができる。(1) 脱落サンプルを除去する。これには、一つでも脱落・無回答があれば消去する**リストワイズ消去法**と分析に必要な変数が脱落している時に標本を消去する**ペアワイズ消去法**に分けられる。(2) 脱落箇所に数値を補完する[15]。これには最終観測値をそのまま延長したり、観測される他の主体の平均値でおきかえたりする**単一値代入法**か脱落箇所を補完した完備データをつくり、それに基づいて推計を行い[16]脱落箇所を完全に埋めた場合に得られるであろうパラメータを推測する**多重代入法**が用いられる。(3) 利用可能データを最大限に生かして分析する。これには、脱落メカニズムを推計して、そのサンプルセレクション・バイアスを調整した上で行動方程式を推計するという **Heckman(1979) の 2 段階推定法**やその拡張、脱落パターンごとの変数の分布と脱落パターンの出現頻度を表す確率分布の積である同時密度関数を推定する**パターン混合モデル**(Pattern-Mixture Model) などがある。

具体的なモデルを考えよう[17]。

---

[15] この方法は統計学者の間でよく用いられ、研究されている。例えば、Little and Rubin (2002) などを参照。

[16] これはさらに (1) 回帰分析法、(2) 傾向スコア法 (Propensity Score 法)、マルコフ連鎖モンテカルロ法（Marcov chain Monte Carlo 法: MCMC）などに分けられる。これらの方法の詳細については岩崎 (2002)、和合 (1998)、坂本 (2004) を参照。先の分類ではランダム脱落 (MAR) を想定していることになる。

[17] 以下の説明は、Fitzgwerald, Gottschalk and Moffitt (1998a, b),Moffit, Fitzgerald and Gottschalk (1999) に依拠している。

$$y = \beta_0 + \beta_1 x + \varepsilon \quad y \text{ は } A = 0 \text{ の時観察可能} \tag{1}$$

$$A^* = \delta_0 + \delta_1 x + \delta_2 z + v \tag{2}$$

$$\begin{aligned} A &= 1 \quad if \ A^* \geqq 0 \\ &= 0 \quad if \ A^* < 0 \end{aligned} \tag{3}$$

ここで A は脱落ダミーで y が観察不可能の時、$A = 1$ となり、観察可能であれば $A = 0$ をとる。

上述のランダム脱落と非ランダム脱落の違いは次のように表現できる。

（i）z と ε は独立しているが、ε と v は独立していない場合、非ランダム脱落あるいは Selection on Unobservables である。

（ii）ε と v は独立しているが、ε と z は独立していない場合、ランダム脱落あるいは Selection on Observables である。

（i）の場合、x と z が全期間観察可能であるとすると、観察可能な y に対する期待値は次のように表せる。

$$\begin{aligned} E(y|x, z, A) &= \beta_0 + \beta_1 x + E(\varepsilon|x, z, v < -\delta_0 - \delta_1 x - \delta_2 z) \\ &= \beta_0 + \beta_1 x + h(-\delta_0 - \delta_1 x - \delta_2 z) \\ &= \beta_0 + \beta_1 x + h'(F(-\delta_0 - \delta_1 x - \delta_2 z)) \end{aligned} \tag{4}$$

ここで h は誤差項 ε と v は説明変数 x と z からはそれぞれ独立していることを表わしたインデックスであり、h' はインデックス h を確率分布に変換したものである。h 関数あるいは h' 関数を特定化し、非線形最小二乗法で推計すれば、(4) 式のパラメータも一致推計が得られる。(4) 式の推計パラメータ β を識別するためには、z と ε が独立で、z のパラメータ δ_2 はゼロではないような制約を課す必要がある。実際このような変数 z を見つけることは難しい。その上、y が観察不可能な場合には、x と z も観察できないことが多く、その場には、上の方法は用いることができない。

（ii）の場合、ε と z が独立していないということは非脱落サンプルに関して (4) 式を最小二乗推計しても一致推計は得られない。z は脱落確率 A に影響を与えるだけでなく、y の条件付き分布にも影響を与える。これは z が y にとって内生変数であることを意味している。すなわち、y の観察可能なデータに基づく分布 $g(y|x, A = 0)$ は、脱落サンプルも含めた全サンプルの分布 $f(y|x)$ とは一致しないのである。

ここで $f(y, z|x)$ を全サンプルに関する y と z の同時分布であるとし、$g(y, z|x, A = 0)$ を条件付き同時分布であるとすると、

$$
\begin{aligned}
g(y,z|x,A) &= \frac{g(y,z,A=0|x)}{\Pr(A=0|x)} \\
&= \frac{\Pr(A=0|y,z,x)f(y,z|x)}{\Pr(A=0|x)} \\
&= \frac{\Pr(A=0|z,x)f(y,z|x)}{\Pr(A=0|x)} \\
&= \frac{f(y,z|x)}{w(z.x)}
\end{aligned} \tag{5}
$$

ここで $w(z,x)$ は次のように定義される。

$$
w(z,x) = \left[\frac{\Pr(A=0|z,x)}{\Pr(A=0|x)}\right]^{-1} \tag{6}
$$

これは脱落しない確率の逆数であり[18]、これを用いると、全サンプルに関する同時分布 $f(y,z|x)$ は次のように表現できる。

$$
f(y,z|x) = w(z,x)g(y,z|x,A=0) \tag{7}
$$

つまり、全サンプルに関する同時分布は脱落しないサンプルの条件付き同時分布を脱落しない確率の逆数でウェイトづけして求めることができる。これを z に関して積分すると、求める全サンプルの確率分布 $f(y|x)$ を求めることができる。

$$
f(y|x) = \int_z g(y,z|x,A=0)w(z,x)dz \tag{8}
$$

つまり、この $w(z,x)$ によってウェイト付けした加重最小二乗法 (WLS) で（1）式を推計すれば一致推定量が得られる。

この結果から、脱落サンプルによるバイアスが無いための条件は以下の通りである。

- z が脱落確率 A とは無関係である場合。この場合は、（6）式のウェイトが 1 になり脱落サンプルの調整は必要がなくなる。
- x という条件の下で y と z が独立である場合。z は内生変数ではなくなり、全サンプルの条件無し分布 $f(y|x)$ は条件付き分布の積分値と一致し、脱落サンプル・バイアスは消滅する。

2.4.3 脱落サンプル・バイアスの検定

では、脱落サンプル・バイアスはどのように検定すればいいだろうか。これにはいくつかの方法が提案されている。

[18] これを inverse probability weight (IPW) と呼ぶ。これは脱落バイアスを取り除くためのウェイトであり、不均一分散を取り除くためにかけるウェイトとは意味がちがう。Wooldridge(2002b, 2004) は IPW に基づいた推計方法の性質を明らかにしている。

最も素朴な方法ではあるが、全サンプルと脱落サンプル、継続サンプルのそれぞれの関心ある変数の平均や標準偏差を比べてみるというのは、脱落サンプルのバイアスの潜在的な大きさを直感する意味では有益である。

(i) 脱落サンプルと継続サンプルの個別変数の平均値の比較。これは、t 検定を行うことによって、2 つの分離したサンプルの分布が統計的に等しいかどうかを検定するものである。具体的には脱落したサンプルが継続していた期間の変数の平均と継続サンプルの変数の平均の差を t 検定することによって、その差の有意性を見ることができる。また継続サンプルの確率分布と脱落サンプルの確率分布が等しいかどうかをカイ二乗検定することもできる。また、似たようなパネルデータ調査があり、その脱落確率に違いがあるとすれば、その外部データの同じような変数を比較することで、脱落サンプルのバイアスの程度に見当をつけることができる。前節で紹介した Newmark and Kawaguchi (2001) などがその例である。

最も初期の統計的検定の提案は Hausman and Wise (1979) であり、それを Nijiman and Verbeck (1992) が拡張したものである。

(ii) 第一の方法は、もともとの標本の一部が何らかの理由で観察不可能になった、あるいは標本に入っているべき対象者のデータが切り捨てられていると考えて、Heckman(1979) の 2 段階推定法を利用するということである。ヘックマンの方法はクロスセクション・データに用いられていたものであり、ここではパネルデータをプーリングして用いる。ヘックマンの方法は多くの統計パッケージに入っており、簡便な方法なので実証的には最も用いられているバイアスの検定方法である。第二の方法は、全サンプルを用いた推計パラメータと継続サンプルを用いた推計パラメータをハウスマン検定（Hausman Test）[19]により比較するというものである。この方法の問題点は、それぞれの推計がもともと一致推計ではないので、ハウスマン検定の検定力が低いと考えられることにある。第三の方法は、ヘックマンの方法に推計バイアスを修正する項をさらにいくつか追加して、その係数がゼロであるかどうかを検定するものである。

ここで論じられた検定方法は、さらに様々に改善された。まず、Fitzgwerald, Gottschalk and Moffitt (1998a, b) , Moffit, Fitzgerald and Gottschalk (1999) は代替案として次の 2 つの方法を提案している。

(iii) 脱落サンプル・バイアスの検定は $A = 0$ という事象に z が説明力を持つかということを見る。

これは、(2) 式を全てのサンプルが揃っていた第 1 期のデータを用いて、最終的に明らかになっている脱落事象がどの程度説明できるかを見る。ここでは被説明変数が $(0, 1)$ の二項選択なのでプロビット分析を行い、z の係数 δ_2 が有意であるかどうかを検定する。

$$A^* = \delta_0 + \delta_1 x + \delta_2 z + v \tag{9}$$

[19] ハウスマン検定については第 3 章で詳しく解説する。

ここで z として y のラグ変数を用いることも可能である。

(iv) 行動方程式モデルを全サンプルについては最小二乗法で推計し、脱落したサンプルの確率を調整した継続サンプルの加重最小二乗法の推計パラメータとを比較し、ハウスマン検定を行う。

これは、（10）式を最小二乗法で推計したパラメータと、

$$y = \beta_0 + \beta_1 x + \varepsilon \tag{10}$$

（11）式を加重最小二乗法で推計したパラメータを比較するものである。

$$y/w = \beta_0(1/w) + \beta_1(x/w) + \varepsilon/w \tag{11}$$

ここで w は（6）式で導いた脱落しない確率の逆数である。この検定は z が内生変数であるかどうかを間接的に検定していることになる。

もう一つの検定として Becketti, Gould, Lillard and Welch (1988) が提案したのは次のようなものである。

(v) 被説明変数 y の第 1 期の値（初期値）y_0 に対して x や第 2 期目以後の A が影響を与えていたかを見ることで、脱落サンプルがバイアスを与えたかどうかを検定しようとするのもである。これは次のような関係から導かれる。すなわち、これまで見てきた（2）式あるいは（9）式の関係を逆にして期待値をとり、ベイズ定理を応用して書き直すと次のようになる。

$$E(y_0|A, X) = \int y_0 f(y_0|x) w(A, y_0, x) dy_0 \tag{12}$$

ここで

$$w(z, x) = \frac{\Pr(A|y_0, x)}{\Pr(A|x)} \tag{13}$$

このウェイトは基本的には（6）式と同じであるが、ここでは $A = 0$ だけではなく、$A = 1$ も含んでいる（A はダミー変数として入っている）。もし、ウェイト w が 1 であれば（12）式において y_0 の条件付き分布は A とは独立となり、以下の回帰式において β_2 は有意ではなくなる。

$$y_0 = \beta_0 + \beta_1 x + \beta_2 A + \varepsilon \tag{14}$$

この係数 β_2 を有意検定することで、バイアスの程度を見ることができる。しかし、この検定は（9）式の検定とペアで考えるべきもので、独立したバイアスの検定ではない。

このように脱落サンプル・バイアスの検定は様々なものが提案されているが、基本的な考え方はクロスセクション・データのセレクション・バイアスの検定問題に基づいており、パネルデータとしてサンプルが逐次脱落していくことに伴うバイアスの発生やサンプルに復帰した場合の取り扱い、サンプ

ルの脱落が特定の時間に集中した場合のバイアスの問題など、まだまだ解決すべき問題は数多く残されている。また、この問題は不完備データの問題とも密接に関連しており、その点に関しては第 3 章で論じたい。

2.4.4 脱落サンプル・バイアスの実証結果

Becketti, Gould, Lillard and Welch (1988) は PSID が時間とともに変容していく姿を確認し、標本抽出した時点から見ると、その標本としての代表制はどの程度維持されているのかという問題を提示した。彼らは前節で提示し（14）式の推計を行い、A の係数が有意ではないことを確認し、脱落サンプルが推計パラメータに有意な差をもたらすことはないと結論づけている。

Fitzgwerald, Gottschalk and Moffitt（1998a）、Moffit, Fitzgerald and Gottschalk (1999) も PSID の脱落サンプル・バイアスを統計的に検定し、（9）式のプロビット分析を行った。その結果、様々な変数を入れて推計を行うと、ほとんどの変数の係数が有意でなくなり、また決定係数も極めて低く、例え、賃金所得の係数が負に有意に効いているとしても、脱落確率そのものを説明する力はほとんどないと報告している。加えて、Becketti, Gould, Lillard and Welch (1988) らが提案した（14）式も推計し、変数にかかる推計パラメータは不変であるが、切片は異なるという仮説が棄却できないことを示した。すなわち、脱落サンプルはパラメータそのものには影響を与えないが、被説明変数のレベルに有意な差をもたらしているということである。

Fitzgwerald, Gottschalk and Moffitt (1998b）では同じ PSID を使い、世代間関係に着目し、脱落サンプルが次世代の経済変数に影響を与えているか、あるいは親子間関係に影響を与えているかどうかを検討した。(9）式のプロビット分析では、脱落確率を有意に説明する変数はほとんど見つからなかった（脱落に対する教育の世代間に及ぶ負の効果は見られた)。また決定係数も低く、分析で用いられた変数で脱落確率を説明する力はほとんどないことがわかった。

Falaris and Peters (1998) は The National Longitudinal Surveys of Labor Market Experience と the Panel Study of Income Dynamics を用いて、学校選択に関して常に回答してきた継続サンプルと脱落サンプルのデータに分離し、最小二乗推計を行い、パラメータの違いを F 検定した。その結果、残留と脱落の違いは推計パラメータにはほとんど影響を与えていないし、影響があるとすれば切片の推計に見られるぐらいであるという結果を導いている。また Heckman の 2 段階推計法を用いて、継続するかどうかの選択モデルを推計し、次いで、教育（年齢）選択モデルを推計すると、2 本の推計式の誤差項の相関 $\rho(Rho)$ が統計的に有意になった。これはサンプルセレクション・バイアスがあることを示唆していると論じている。

Zabel (1998) は PSID と the Survey of Income and Program Participation (SIPP) の脱落率を比較する目的で（9）式のロジット分析を行ったが、初年度をどの時点に設定するかによって結果が違ってくることを示している。す

なわち、SIPP1984 と SIPP1990 では脱落のプロセスが違うことが示唆されている。Zabel (1989) は Heckman の 2 段階推計法を拡張した労働供給モデルも推計したが脱落サンプルが推計パラメータに与えた影響は小さいと報告している。

Burkam and Lee (1998) は High School and Beyond Study (HS&B) について、学業成績を被説明変数とする（14）式のようなモデルを推計し、脱落サンプルの効果を調べたが、もともと脱落率が低い調査ではあるが、脱落サンプル・バイアスを調整せずにデータを用いると人種による成績への負の効果を過大評価してしまうおそれがあると報告している。しかし、このバイアスは特定の人種やグループによってもたらされているものではないことも確認されている。

日本の『消費生活に関するパネル調査』の脱落サンプル問題に関しては坂本（2003、2004）が実証分析を行っている。

脱落ダミー（脱落＝１、継続＝０）を被説明変数、前年の個人属性・変数(「学歴」「都市規模」「世代」「子供の人数」「病気・事故・災害」「親との同居」「年収」「有業・無業」「結婚が決まっている」「新婚」「世帯員が死亡」「出産」「生活時間」「夫婦関係満足度」）を説明変数とする（9）式のプロビット分析を行っている。既述のコーホート A の内、無配偶についてみると、「結婚予定」ダミーが全てのケースで有意に効いている。本人年収は高いほど脱落率が下がる。「仕事時間」が短く、「趣味・娯楽時間」が長いほど脱落確率は低い。有配偶サンプルでは若い「世代」ほど脱落率が高く、「新婚」ダミーは脱落率を高めることがわかった。経済状況は無配偶サンプルとは逆に本人収入、夫の収入が高いほど脱落率が高くなっている。コーホート B の無配偶サンプルでは「結婚予定」が脱落確率を上昇させ、「家事・育児時間」の増加はむしろ脱落確率を低下させている。有配偶サンプルでは若い「世代」ほど脱落確率が高くなっている。日本のデータでは「結婚」を機に調査から脱落する傾向があることが確認されている。無配偶では低学歴、低収入ほど脱落確率が高いというアメリカの結果と合致しているが、有配偶では逆の結果になった。

また、ヘックマンの 2 段階推計によるサンプルセレクション・バイアスの修正を施した推計と施さない推計のパラメータをハウスマン検定を行うことで比較した。ここで用いたモデルは世帯消費を世帯主の年齢、学歴、子供の数、長子の年齢などで説明するものである[20]。検定結果はパラメータが有意に違うという帰無仮説は棄却され、脱落サンプルによるバイアスは大きくないことが確認された。

途上国のパネルデータの脱落サンプル・バイアスを計量経済学的に分析した研究には、前述の Alderman, Behrman, Kohler, Maluccio and Cotts Watkins (2000) と Maluccio（2000）がある。まず Alderman,*et al.*(2000) では（9）式のプロビット分析と（14）式の最小二乗推計を行ったが、ボリビア、ケニア、南アフリカではいずれも有意な脱落サンプル・バイアスは発見されなかった。

[20] これは Nijiman and Verbeek (1992) のモデルと検定方法を踏襲している。

Maluccio（2000）は Alderman,*et al.*(2000) と同じ南アフリカのデータを使って基本的には同様の検定を家計支出モデルについて行い、脱落サンプル・バイアスが見出されると報告している。またヘックマンのセレクション・モデルを用いてセレクション・バイアスを修正すると、ハウスマン検定によってパラメータが等しいという仮説も棄却されている。これらの結果から、脱落サンプル・バイアス検定はモデルに強く依存していることがわかってきた。

これまでの実証結果を見る限り、脱落サンプル・バイアスは少なくとも統計的には大きくないという結果が多い。しかし、同時にそれは計量モデルに依存しており、検定方法自体も様々な脱落パターンに対応しておらず、かなり限定的な制約下で行われていることに注意しなければならない。PSID のように元のサンプルの 50% 以上が脱落していてもバイアスは見られないという検定結果には大いに留保が必要であると思われる。しかしながら、脱落サンプル問題があるからといってパネルデータを使うことを回避するのは建設的ではない。むしろ、そのバイアスの大きさを常に統計的に確認した上でデータを用いる注意深さが要求されているのだと考えるべきであろう。

2.5 データセットの作り方

パネルデータ調査の結果集められたデータをどのように保存し、データベースを構築していくかということは重要な問題である。もともとパネルデータ統計として設計された調査であれば、個々の調査主体の ID（認識番号）は固定されており、同じ質問項目に対しては同一の変数名、同一の選択肢順序、データベース上の配置 (location) を同一とすることで複数年の調査をマッチングすることが大幅に容易になる。逆に、ある調査だけ年度だけ特別に行った質問項目はデータベース上の配置は、他の年度の調査の他の質問項目の配置と重ならないように配慮すべきである。

もともとがクロスセクション統計であり、パネル化することが意識されていない調査では、個々の調査主体の ID は固定されていたとしても、質問項目が都市によって微妙に変化したり、変数名が変更されていたり、データ-ベース上の配置に至っては過去のデータベースとの整合性はほとんど配慮されていないことが多い。この場合、複数年のデータ-ベースをマッチングさせてパネルデータ化するには、根気強い調整が必要になる。この部分で慎重に作業を行わないと、後でいかに高度な手法で推計しようが、実証結果の信頼性は低くなる。これまでの経験からすると、この作業を通してデータの性質もわかってくるので、ここにかなりの時間とエネルギーを用いるべきである。

2.5.1 データセットの構造

パネルデータを実証で用いる場合には次のような構造をしていることが多い。

図表 2.1　パネルデータセットの構造

パネルデータにおいて最も特徴的な変数は ID(個体認識) と Time(年月) であり、一般に、統計ソフトはこの 2 つの変数を明示することでパネルデータであると認識する。

STATA においてはパネルデータ分析のはじめに次のようなコマンドを入れる必要がある。

iis ID (varname_i)

tis Time (varname_t)

あるいは

tsset ID Time

とコマンドを入れることで時系列データであると同時に ID も違うデータであると認識する。

ラグの導入

第 4 章で詳しく解説するが、パネルデータの変数のラグをとって、ダイナミック・パネルデータとして分析する手法が多用されるようになってきている。パネルデータの場合、単純な時系列データと違い、異なった経済主体 (ID) のデータが積み重なるようなかたちでデータを構成しているので、単純に時系列データのラグの入れ方を用いると経済主体の時系列上最後の観察データが次の経済主体の時系列の最初の観察データとして入ってくる。そのような問題を回避するためには、このようなラグを導入するを行う前に上述したようなパネルデータであることをプログラムに認識させるコマンドを書く必要がある。間違いのないパネルデータにおけるラグの作り方は次のようなコマンドを書くとよい。

```
gen  Y_1=Y[_n-1] if ID[_n-1]==ID[_n] &
     Time[_n-1]+1==time[_n-1]
```

これは Y の 1 期ラグとして Y_1 を作るときに同じ ID の主体に対してのみ時系列ラグと入れるように指定した条件をつけたものである。

時間表示

時間 (Time) を表す変数は、観察間隔が一定であることが望ましいが、必ずしも物理的時間が一定である必要はない。例えば、実際の調査が特定の年月日に行われて、データベースの Time の項に年月日まで記入されている場合があるが、分析上大切なことが、ある年度に行われた何回目の調査であるということであれば、年月日が個別主体ごとに異なっていたり、調査と調査の間隔が一定でないことになり、パネルデータとして混乱が生じる可能性がある。このような場合には、時間の変数を連続した自然数に置換した方がいい。例えば、次のようなコマンドを書く。

replace time=1 if time==20010811
replace time=2 if time==20010915
replace time =3 if time==20011009

時間に関しては時間ダミーを入れてその時間に各経済主体に共通に与えたマクロ・ショックを推計することがある。その場合には次のようなコマンドを用いるとよい。

gen Aug2001=1 if time==1
replace Aug2001=0 if time !=1
gen Sep2001=1 if time==2
replace Sep2001=0 if time !=2
gen Oct2001=1 if time==3
replace Oct2001=0 if time !=3

属性ダミーの作成

属性データとしては、企業データであれば、業種コード、所在地都道府県コード、設立形態、設立年などが入っているし、家計データであれば住所の都道府県コード、世帯区分、世帯人員、就業人員、住居の所有関係、世帯主年齢などが入ってくる。これらの変数には設立年や世帯人員のようにそのままの数字が入っている場合はいいが、項目によっては調査側が慣例として用いている区分番号が割り振られていることがある。例えば、世帯区分が１．勤労世帯、２．個人営業世帯、３．その他世帯とある場合、この数字の 1，２，３には何ら経済的な意味はない。このような場合には勤労世帯を１、それ以外を 0 とするような勤労世帯ダミー、個人営業世帯を１、それ以外を 0 とするような個人営業世帯ダミーを作ることが望ましい。住居の所有関係から例を作ると次のようなコマンドになる。

gen ownhouse=1	if	juukyo==1
replace ownhouse=0	if	juukyo !=1
gen privrent=1	if	juukyo==2
replace privrent=0	if	juukyo !=2
⋮		⋮
gen koudan=1	if	juukyo==5
replace koudan=0	if	juukyo !=5

全く同様に産業コードも次のようにダミー化できる。

gen mining=1	if	sangyou==1
replace mining=0	if	sangyou !=1
gen construct=1	if	sangyou==2
replace construct=0	if	sangyou !=2
⋮		⋮
gen retail=1	if	sangyou==6
replace retail=0	if	sangyou !=6

経済変数の外れ値の処理

一般に、経済変数 Y が実数で表示されていれば、それはダミーで表示されたカテゴリーデータより多くの情報を含んでおり、その情報をできるだけ大切に利用して統計分析を行うことが望ましい。しかし、以下に貴重な経済変数とはいえ、必ずしもすべてを利用すべきであるということではない。ある経済主体は無難な質問には答えるが、分析上関心のある質問には答えていないとすれば、その経済主体を分析から外しても情報のロスにはならない。ある経済モデルに基いて統計データを分析しようとすると、明らかにその経済モデルが前提としている経済行動にそぐわない主体が存在することに気づく。これは統計上外れ値として知られているが、桁外れの収益を上げている企業や資産家を平均的な経済主体として扱うと、統計的にそれらの外れ値が結果を歪めてしまうことがある。一般に、ミクロ計量経済学では、このような外れ値を除外して分析することが多い。外れ値の処理に関して一般的なルールがあるわけではないが、関心のあるいくつかの変数の平均から ±3× 標準偏差を外れるサンプルは標本全体の 0.3 %以下であり、平均から ±4× 標準偏差を外れるサンプルは標本全体の０ 0.1% 以下であることが知られている。われわれは通常サンプルをできるだけ無駄にしないという目的で平均から ±4× 標準偏差を外れるサンプルを除外している。

経済変数のカテゴリー化

経済変数の中には、はじめに二項選択によって（購入、非購入）あるいは（参加、不参加）を決め、次に購入を決めた場合にいくら購入するか、参加と決めた場合にどのように参加するかを考えるタイプの変数が含まれている。はじめの二項選択をダミー変数として表現して経済主体をカテゴリー化することが有益なケースがある。例えば、企業であれ家計であれ負債がある主体とない主体に分けて、負債のある主体の経済行動と負債のない主体の経済行動を比較することがある。負債には短期負債と長期負債があり、長期負債について考察する場合には調査期間中に長期負債を負えばその期間中は常に負債を負っているというカテゴリーに入る。この行動ダミーは次のように作ることができる。

```
gen debtdum=1      if    kariire>0
replace debtdum=0  if    kariire==0
```

しかしこのカテゴリー化では長期負債を調査期間中に返済して無負債になった主体は debtdum=1 から debtdum=0 に途中でシフトすることになる。経済分析の関心が負債のある主体とない主体の違いにあるのであれば、このカテゴリー化で問題はないが、はじめから負債のなかった主体と負債を返済して無負債になった主体では経済行動に違いがあるはずだという観点に立てば、ダミーを増やして（1）調査期間中一貫して無負債の主体、（2）調査期間中、無負債からスタートし、途中で負債を負いそのまま継続している主体、（3）

調査期間中、はじめは負債があったか、途中で負債を返済し、その後、無負債を続けている主体、（4）調査期間中負債と無負債を繰り返している主体あるいは（１）〜（3）のいずれにも該当しない主体と分けることも考えられる。

ダミーを用いて経済主体をカテゴリー化してやる方法は医学や社会学、心理学の分野では多用されているが、経済学ではそれほど後半に使われてはいない。しかし経済パネルデータ分析の一つの方向性としてはカテゴリがデータ分析の手法の応用が重要になってくるものと思われる。とりわけパネルデータを用いて政策プログラムの評価を行う場合には処理グループと管理（非処理）グループに分離して比較検討することが必要になる。その際カテゴリカルデータ分析に手法を用いざるを得なくなる。

経済変数のダイナミックなカテゴリー化

パネルデータにおける経済変数は時間とともに変化しうるので、初期条件やある特定の時点でのカテゴリー化だけでは分析の目的にそぐわない場合がある。例えば、企業収益は常に変動しているが、企業行先に問題があると見られるかは収益が二期続けてマイナスの場合であるとすると、そのような企業をカテゴリー化するためには調査期間中に収益が二回連続して負になった企業をサンプルから選び出す必要がある。そのようなカテゴリーを”loss”として表現するコマンドは次のように書ける。

```
sort ID year
gen flg1=1 if ID==ID[_n-1]&year==year[_n-1]+1&profit<0&profit[_n-1]<0
by ID, sort: gen flg2=sum(flg1)
by ID: replace flg2=flg2[_N]
gen loss=0
replace loss=1 if flag2>0
```

収益が一度も負になったことのない健全な企業を”good”というカテゴリーで表す。

```
gen flag3=1 if profit>=0 | profit==.
by ID, sort: gen flg4=sum(flg3)
by ID: replace flg4=flg4[_N]
gen good=0
by ID, sort: replace good=1 if flag4==_N
```

収益が負になることはあるが二期続けて負にはなったことのない企業は”good”でも”loss”でもないということで”alive”というカテゴリーに入れる。

```
gen alive=0
replace alive=1 if loss==0&good==0
```

これで調査期間中すべての企業を三つのカテゴリーに分類することができる。

経済変数のカテゴリー化はそれぞれの研究内容、データに利用可能性に応じて決まってくるが、論理的に厳密にカテゴリー化されていれば、STATA でプログラムを書くことは容易である。

図表 2.1　パネルデータセットの構造

ID（個体認識）	Time（年月）	Characteristics（属性）	Variable（度数）
1	1980	X_1	Y^1_{1980}
1	1981	X_1	Y^1_{1981}
1	1982	X_1	Y^1_{1982}
1	1983	X_1	Y^1_{1983}
1	1984	X_1	Y^1_{1984}
1	1985	X_1	Y^1_{1985}
2	1980	X_2	Y^2_{1980}
2	1981	X_2	Y^2_{1981}
2	1982	X_2	Y^2_{1982}
2	1983	X_2	Y^2_{1983}
2	1984	X_2	Y^2_{1984}
2	1985	X_2	Y^2_{1985}
3	1980	X_3	Y^3_{1980}
3	1981	X_3	Y^3_{1981}
⋮	⋮	⋮	⋮
⋮	⋮	⋮	⋮