あるがゆえに、より多くのことを知ることができるという、その情報量である。第二は、事実をより正確に、偏りを少なく推計することを可能にするという面である。

サンプルが替わらないパネルデータでしかわからないことは、第一に、個々人の変化の状況である。それは、状態の変化、個人間の順位の入れ替わりなどである。第二に、量的にどの程度の変化をした人がどのくらいいるのかという、変化率の分布のような情報も、パネルデータでしかわからない。第三に、どういう人がどういう変化をしているのか、あるいは、どう変わりやすいかがわかるし、逆に、こういう変化をしている人はもともとどういう人かという情報もとることができる。

(3) パネルデータの利点②（入門パネルデータによる経済分析③）

複数の同一主体を追跡していくパネルデータは、調査回数を重ね、長時間にわたる情報が得られるようになって、その価値を増していく。

第一に、単純な事実として、ある属性の人たち、あるいはある行動をとった人たちがその後どうなっていったかがわかる。第二に、長期間の観察を行うと、さまざまな変化が、長期的な変化、恒久的な変化であったのか、それとも短期的、一時的な変化に過ぎず、すぐに元に戻るようなものであったのかを区別できる。第三に、長期間追跡することによってわかったことから、そのわかった人たちはどういう人たちかという分析も可能になる。すなわち、長期の観察による類型化である。

(4) パネルデータの利点③（入門パネルデータによる経済分析④）

同一の個人、企業を追跡していくパネルデータは、時系列データ（タイムシリーズ・データ）と横断面データ（クロスセクション・データ）の両方の性質を兼ね備えている。時系列的性質からは、個人等の変化の情報が得られる。横断面的性質からは、個人間の違いに関する情報が得られる。それらの情報を組み合わせることによって、時系列データや横断面データだけではわからない、より複雑な動きをとらえることができる。例えば、時系列データは変化をとらえるが、パネルデータでは、それに複数の個人を追うという横断面的な情報が加わることによって、その変化をいくつかの要因に分解することなどが可能になる。

また、パネルデータは個々人が時間を追って変化していく様子を追跡していくものである。その変化について、「年齢効果」「時代効果」「世代効果」の３つの効果をとらえるという「コーホート分析」がよく行われる。こうした３つの効果に分けてとらえることは、横断面データや時系列データでは不可能である。ただ、このコーホート分析では、化悪事代、各年齢（したがって各世代）のデータがそろっていても、３つの効果を一意に定めることができないという「識別問題」がある。

(5) パネルデータの利点④（入門パネルデータによる経済分析⑤）

これまでパネルデータの利点として、さまざまな変化の態様を追うことができることを論じたが、ここでは、これらの変化に関する情報を組み合わせるなどして、どういった分析が可能になるかを述べる。

第一に、生活のある面で変化が起こった時に、他の面ではどのような変化があるのかということをみることができる。例えば、大きなライフイベントである結婚や出産、離婚によって、所得や消費などの暮らしぶり、心理状態などがどう変わるのかを直接みることができる。

第二に、パネルデータでは、さまざまな変数間の因果関係の識別やその強さの計測を容易にする。こうしたことが政策効果の分析や、個人の選択、行動メカニズムの解明の可能性を広げる。

まず、横断面データ（クロスセクション・データ）に対する利点として、①時間的前後関係の情報を提供することによって、因果関係の識別を容易にする。Xから Yへの因果関係が存在するかどうかを検討する際には、一般に（a）両者の動きの間に相関があること、(b)見せかけの相関でないこと、(c)時間的前後関係（Xが Yに先行するか、同時であること）という 3 つの基準を満たしているかどうかを調べる必要がある。そして、②横断面のサンプルでは区別できないことの影響の計測である。横断面では、原因となる変数の値がすべてのサンプルで同一の場合があり、その影響が個人の行動にどの程度影響するかは、時間をかけてみていくしかない（この点に関しては、時系列データであればわかることで必ずしもパネルデータでなければならないものではない）。

次に、時系列データ（タイムシリーズ・データ）に対する利点として、①複数の主体間での比較、②時間を通じて換わらないことの影響の計測が、パネルデータでは可能となる。

第三に、パネルデータでは、個人の周囲で何かが変わったような場合、その変化によって個人がどれほど影響を受けるか、新たな政策が発動されて、それが個人の行動や状態にどれほど影響するかという政策効果の測定が可能になる。パネルデータはその情報量が多いことによって、サンプルの無作為性（ランダムネス）を確保しやすくする。アプローチとしては、Before and After アプローチと、Differences-in-Differences アプローチがある。

Before and After アプローチは、同一の人について、政策が発動された前後でその変化を比較する方法である。Fixed Effects（固定効果）アプローチと呼ばれることもある。このアプローチは、当該政策が発動されていなければ、その個人には変化がなかったという前提に立ち、政策発動後の実際の変化をもって政策の効果とみなすものである。しかし、マクロ変動の影響など、個人に影響する他の要因もありうる。それらの影響が取り除かれないと、このアプローチでは必ずしも適切な影響把握ができないことになる。

そこで、同じようにマクロ変動を受けている人たちの中で、政策の対象となった人とならなかった人とで、政策発動前後の変化を比較することによりマクロ変動の影響を除いてみようというのが、Differences-in-Differences アプローチである。政策の対象になった人

と、対象にならなかった人との変化の差を政策の効果とみなすものである。このアプローチは、マクロ変動の影響が、政策の対象になった人とならなかった人との間で違わないという前提に立っている。その上で、政策発動前からあった違いをコントロールし、無作為に抽出した集団同士を比較するかのような状況を作りだしている。この手法は純粋な実験ではないが、実験計画の考え方を取り入れ、政策の対象となった人は処置群(treatment group)、政策の対象とならなかった人は対照部または比較群、統制群(control group)と呼んでいる。

なお、横断面データ（クロスセクション・データ）しかない場合も、例えば、政策の行われた地域とそうでない地域の状況を比較するということも考えられる。しかし、その場合は、それらの地域間で政策発動前から違っていたかどうかを知ることができない。政策発動前は同じであったとみなすことも考えられるが、それでは無理がある場合も少なくなく、計測値にバイアスが入り込む余地もある。

（相馬直子）

稲葉昭英著
「Pooled time series モデル」
『家族社会学研究』14-1, pp.5-10（2002年）

本論文は、アメリカの家族研究において近年パネルデータが重用されるようになった事情をふまえ、パネル調査の特性を生かした統計解析法について解説したものである。Pooled time series モデルの解説となっているが、具体的にはランダム効果モデルおよび固定効果モデルについて、モデルの違いや実証分析に用いる際の注意点などが述べられている。

固定効果モデルとランダム効果モデルに共通な基本のモデルは以下のである。

$$y_{it} = \alpha_i + \beta' x_{it} + \varepsilon_{it} \qquad i = 1,2,\cdots n, \qquad t = 1,2,\cdots T$$

このようなモデルのパラメターを推定する際に、固定効果モデルでは LSDV(Least Squares Dummy Variable)推定量を求め、ランダム効果モデルでは GLS(Generalized Least Squares)推定量を求めることになる。

(1) 固定効果モデルにおける LSDV 推定量

LSDV 推定量を得るための３つの方法を紹介。ひとつめは、個人効果αの推定にダミー変数を用いる方法。しかし計算量が多くなるので、代わって Within 推定量と Between 推定量が考案されている。Within 推定量を得るためのモデルは、

$$y_{it} - \bar{y}_{i\cdot} = \beta'(x_{it} - \bar{x}_{i\cdot}) + \varepsilon_{it} - \bar{\varepsilon}_{i\cdot}$$

と表され、説明変数、被説明変数、撹乱項いずれも個人内の平均からの偏差で表現される。このモデルでは、期間を通じて一定の変数（属性変数など）は入れられない。また、平均値の大小が結果に影響しない。一方 Between 推定量を得るためのモデルは、

$$\bar{y}_{i\cdot} = \alpha + \beta' \bar{x}_{it} + \bar{\varepsilon}_{it}$$

と表記される。これは個人の平均値間の線形モデルと言える。つまり α は個人間で共通な定数項であり、個人特性は撹乱項に含まれる。

(2) ランダム効果モデルにおける GLS 推定量

GLS 推定量は Within 推定量と Between 推定量の加重平均である。このモデルは最初の式を以下のように書き換える。

$$y_{it} = \alpha_i + \beta' x_{it} + u_i + \varepsilon_{it}$$

u_i は i 番目の観察対象に固有で時間的に一定の個人効果となる。ただし u_i と x_{it} は独立である（u_i はランダムである）という仮定が置かれる。GLS 推定量は個人効果が大きいほど Between 推定量の比重が小さくなるようなパラメターでモデル化されている。

後半では、二つのモデルの使い方について解説されている。ランダム効果モデルは、個人

内の効果、個人間の効果双方を推定に用い、時点間で変化しない属性変数もモデルに投入できる利点がある一方で、個人効果が説明変数と独立であるという強い仮定が置かれる。社会科学では社会的属性が個人特性に関連があるのは自明であるので、適用範囲が限定される。

一方で、固定効果モデルは個人効果が説明変数と独立であろうとなかろうと不偏推定量が得られる。ただし、時間で変化しない変数（の主効果）はモデルに投入できない。また、個人内偏差は平均値の大小にかかわらずその絶対量の重みが等しいという仮定に対処するためにも工夫が必要である。たとえば、個人間の平均値の分散が大きくて、平均からの偏差の実質的意味が平均値の大小によって異なる場合（TOEFL500 点からの 20 点上昇と、600 点からの 20 点上昇の違い）、被説明変数が同質になるよう標本を分割し、改めて固定効果モデルを適用する方法などが有効である。

本稿では、標準的な手順として、ランダム効果モデルが適用可能かを検討し、統計的前提が満たされない場合は固定効果モデルを適用するということが勧められている。個人効果が説明変数と独立であるという帰無仮説の検証には、ハウスマン検定を用いる。

（補足）SAS を用いたランダム効果の推定

以下では、ランダム効果モデルを SAS を使って推定する場合のプログラムについて、概説する。分析例は SAS Online Document Version Eight で紹介されているものである。

Milliken and Johnson (1984) が示した不均衡混合モデルの例では、固定効果とされる 3 種の機械とランダム効果とされる 6 人の雇用者が研究対象である。各雇用者(person)は、各機械(machine)を、別の時期に 1 ～ 3 回操作する。従属変数は生産品の量と質を評価した全般的な得点(rating)である。

データ（データ名：machine）は以下のような内容になっている。

```
data machine;
   input machine person rating @@;
   datalines;
1 1 52.0  1 2 51.8  1 2 52.8  1 3 60.0  1 4 51.1  1 4 52.3
1 5 50.9  1 5 51.8  1 5 51.4  1 6 46.4  1 6 44.8  1 6 49.2
2 1 64.0  2 2 59.7  2 2 60.0  2 2 59.0  2 3 68.6  2 3 65.8
2 4 63.2  2 4 62.8  2 4 62.2  2 5 64.8  2 5 65.0  2 6 43.7
2 6 44.2  2 6 43.0  3 1 67.5  3 1 67.2  3 1 66.9  3 2 61.5
```

```
3 2 61.7   3 2 62.3   3 3 70.8   3 3 70.6   3 3 71.0   3 4 64.1
3 4 66.2   3 4 64.0   3 5 72.1   3 5 72.0   3 5 71.1   3 6 62.0
3 6 61.4   3 6 60.5
;
```

以下が、混合モデルの例である。machine*person など、ランダム効果が含まれる交互作用項は、ランダム効果として指定される。

```
proc glm data=machine;
    class machine person;
    model rating=machine person machine*person;
    random person machine*person / test;
run;
```

RANDOMステートメントにおけるTESTオプションは、GLMプロシジャにおいて、personと machine*person をランダム効果とした場合に基づくF検定を行う。

なお、混合モデルはMIXEDプロシジャによっても推定できる。

```
proc mixed data=machine method=type3;
    class machine person;
    model rating = machine;
    random person machine*person;
run;
```

（岩澤美帆）

山口一男著
「パネルデータの長所とその分析方法：常識の誤りについて」
『季刊家計経済研究』62, pp. 50-58（2004年）

本論文は、パネルデータに関するいくつかの「神話」「誤った常識」「広範に存在する不十分な理解」について、その誤りをいくつか指摘することで、パネルデータ分析の「深さ」について論じられている。それは大きく以下のように整理できる。

(1) パネル調査の長所について（神話１・２）

神話１．パネルデータは、マクロな時系列的変化を分析するのに優れている。
神話２．パネルデータは、例えば転職などのイベント X が収入などの従属変数 Y の変化にどう影響するかを見るのに適しているが、これはパネル調査をしなくても一回調査で転職者について前職の収入を調査すれば同様の変化の情報が得られる。

神話 1 について、マクロな時系列分析ならば、独立な標本の繰り返し調査(repeated cross-sectional survey)の方がパネル調査より優れている。パネル調査の真の利点は、次の点にある。すなわち、マクロな時系列変化そのものでなく、変化をミクロな個人のレベルでとらえることができる点。共変動する２変数 Y と Z で、どちらがどちらに影響を与えているか、少なくとも２時点の観察値があれば、相互的影響の同時推定モデル cross-variable、time-lagged effects を調べることで分析ができる点。態度や意識などの変数もその持続性や安定度についての個人差の情報が得られ、変化の予測に対して極めて有利な点である。

神話 2 について、比較のためのデータは一回調査でも回顧によって得られるが、正確さの点でパネル調査の方が明らかに優れている。

(2) X と Y_{t-1} の交互作用効果の利用について（神話３・４）

神話３．X の Y の変化への影響について、回帰分析では、変化の動きの方向を区別して推定できない。
神話４．態度や意識 Y の安定性が例えば教育レベルなどの X に依存するか否かは、Y_t の予測において Y_{t-1} と X の交互作用を見ればよい。

神話 3 について、線形の Y の場合、Y を順序のついたカテゴリーであらわし、ロジスティック回帰（二分法の時）か累積ロジット(cumulative logit)回帰分析で行えば容易に区別できる。一般に、X が Y の上方方向の動きを促進（減少）させるのか、下方方向の動きを

減少（促進）させるのかは、単に Xが Yの変化に正または負に影響するという以上に、一歩変化のメカニズムに踏み込んでおり、理論的に重要であり、パネルデータの利用によりこうした分析が可能となる。

神話４について、態度や意識 Yの安定性が Xに依存するか否かは Y_tの予測において Y_{t-1}と Xの交互作用効果を見ればよいという理解でよいかという点だが、態度や意識の安定性は Y_{t-1}の Y_tへの影響の程度の異質性だけでは適切に計れない可能性が大である。個人の態度や意識の安定度を潜在変数 Zで表し、Yの回帰モデルと Zの回帰モデルの同時モデルを応用するといった分析も考えられる。

(3) 因果分析に対するパネルデータの貢献について（神話５、６、７）

神話５．時間とともに変化するイベント（例えば転職や結婚など）Xの（収入や健康など）Yへの影響を見るとき、選択バイアスとは、Xを経験したグループと経験しなかったグループについて、Xの経験以前に２つのグループ間に存在していた個人差により生じる Yのグループ差のことをいう。

神話６－１．パネルデータによる回帰分析で Y_tの予測を、Y_{t-1}を制御して（説明変数に加えて）行えば、Yの変化の予測の分析をすることになる。

神話６－２．パネルデータで Y（例えば収入や健康）の変動について、２時点間で起こったイベント X（転職、離婚など）の影響を計るのに、イベントを経験することになる者と経験しない者との間にすでに時点 t-1 で存在していた個人差の影響を排除してXの影響を見るには、Y_tを予測する回帰分析で Y_{t-1}を制御して（Y_{t-1}を Y_tの説明変数に加えて）Xの影響を見ればよい。

神話７．$Y-Y_{t-1}$を従属変数とする回帰分析は、観察されない個人の異質性を制御できる点で長所があるが、Y_tの Y_{t-1}からの独立を仮定する上に「平均値への回帰(regression to the mean)」の問題もあるので利用すべきでない。

神話５は、誤りとはいえないが不十分な点がある。例えば、離婚や転職がそれを実際に経験した者に不利益をもたらしたかということの過去の評価は得られても、離婚や転職はもし経験すれば不利益をもたらすかどうかという、いまだ実現せずかつ経験したグループへの選択メカニズムが異なる場合への答えはデータから得にくい。また、同様の理由で、統計的因果分析は、実際に行われた政策が成功したかどうかを評価できるが、これから採用される、特に政策に影響される人々の選択メカニズムが異なるような政策が成功するかどうかは判断できないことが多いということも意味する。

神話６－１は完全に誤りであり、より厳密に述べた神話６－２も同様に誤りである。「Yt

の予測に Y_{t-1}を制御する」ということは「Yの変化を説明しようとする」ことでは全くなく、「時点 t-1 での Yの個人差の影響を除外すること」とも異なる。Y_{t-1}を制御することは「時点 t-1」における Yの違いがあり、かつ Y_{t-1}が Yt に影響するという、その二つのことの組み合わせから起こる、時点 t-1 の Yの差が時点 t の Yの差として生じる持ちこみ効果を除外する」ことを意味している。したがって、Y_{t-1}が Y_tに全く影響しなければ、いくら時点 t-1 で Yに差があろうと時点 tに持ち込まれる差は 0 となり、時点 tで Yに差があれば、別の理由に帰せられる。

また、「時点 t-1 で存在している個人差の影響を排除して Xの影響を見る」という表現には暗黙のうちに「時点 t-1 と tの間で Xの変化を経験することになるグループと Xの変化を経験しないことになるグループ間の事前の差」の制御という含意がある。すなわち「観察されない(あるいは制御されない)個人差」があり、それが Y_{t-1}に影響を与えている場合、その影響も含めて排除するという含意であるが、単に Y_{t-1}を独立変数として制御したのでは、Yの持込効果だけの除外となるため、そういったより一般的な個人差の影響は全く除外できない。より一般的な個人差の影響の除外の問題は、以下の、Xの状態への選択バイアスと、それを取り除こうとする fixed effects model の利用に関係している。

神話 7 について、「平均値への回帰」は、Y_{t-1}が $\Delta = Y_t - Y_{t-1}$に影響を与えると「問題が起こる」という議論だが、実際に問題なのはパラメターの推定値の一致性(consistency)で、そこで問題になるのは Y_{t-1}の Y_tへの影響であり、それがなければ、当然 Y_{t-1}と Δ_tの相関係数は－1 で Y_{t-1}は Δ_tに強く影響するが、これは全く問題が起きない。

ただし、$Y_t - Y_{t-1}$について注意を要するのは、それを回帰分析の従属変数として用いるときであり、Y_{t-1}が Y_tに影響すると仮定するか否かに大きく依存する。

差分を用いる回帰分析は、fixed-effects model に現れる。このモデルの長所は、時間と共に変化する説明変数について（時間によって変化しない）個人差に基づく状態への選択バイアスを完全に取り除くことができる点である。したがって、Xの変化の影響は、その状態変化の経験者の間での Xの Yへの因果的影響を表すと解釈できる。一方、このモデルの短所は、Yの観察時に時間とともに変化しない変数 Xの影響は測定不能である。Yの変化のリスクのある時間以外で変化する Xの因果的影響は、fixed-effects model では計れない。

Y_{t-1}の Y_tへの影響がある時は、fixed-effects model に問題が起こる。Time-lagged effects で、Y_tが Y_{t-1}や Y_{t-2}だけでなく Y_{t-3}や Y_{t-4}にも依存するなどの場合は、従属変数が離散的な変数の場合、状態継続時間依存をより一般的に取り扱うイベントヒストリーモデルを用いるべきである。ただし、Y_{t-1}の Y_tへの影響がある時、fixed-effects model の利用に注意が必要であるという意味であり、「不可能」では全くなく、利用条件が満たされれば強力な分析方法である。

因果分析上、fixed-effects models について特に留意すべきことは、このモデルは Xの値の変化を経験した者についての個人内の Xと Yの関係から Xの効果を測定しているので、

あくまで経験したグループ内の X の効果であって、全体での平均的効果ではないという点である。

さらに、回帰モデルでの「観察されない異質性」の制御には fixed-effects model のほかに、random effects model がある。これは観察されない異質性を一定の分布を持つランダム変数で表すが、このモデルは Y_{t-1} は Y_t に対する影響の過大評価を修正する機能があり、また時間で変化しない説明変数をモデルで用いることができるという利点もあるが、その異質性のランダム変数は、X との独立を仮定しているため、状態 X への選択バイアスは取り除けないという問題がある。

しかし、random effects models を拡張して従属変数 Y とそれとの因果関係が問題になる特定の説明変数 X の決定の同時モデルを考え、その誤差項間の相関の有無によって結果が異なるかどうか見る方法や、bivariate probit モデル等の方法も開発されている。ただし、理論的活経験的に選択プロセスの適切な知識が必要とされ、fixed effects model より知識の要求度の高いモデルである。

このように、パネルデータは分析を複雑にもしたが、因果関係の解明など、他の調査データでは不可能な分析を可能にし、社会・経済の実態の解明や予測、政策評価などへの利用を大きく前進させたといえる。

（相馬直子）

北村行伸著
「第２章　パネルデータの調査方法と構造」
『パネルデータ分析』, pp. 27-56（2005 年）

本書はパネルデータの分析手法およびそれを主に経済学に応用した研究を紹介することを目的として書かれている。その中で第２章は、パネル調査の実施およびデータセットの作り方について書かれているので取り上げたい。

パネルデータの調査方法上の問題としては、調査対象の選択範囲（coverage）、非回答(non-response)、脱落サンプル(attrition)が重要と指摘されている。パネル調査の特徴、主要な問題点および解決方法として挙げられていることを列記する。

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、(1)標本設定時脱落による歪み、(2)継続時脱落による歪み、(3)調査慣れがもたらす歪み、(4)回答者の同一性の確認、回答誤記(5)のレベルで検討されるべきであるとされる。脱落サンプルについては、一般的に社会経済的地位の低い人に多い傾向があるが、日本の家計研の調査では、有業高所得者、結婚を理由にした脱落が多いといった特徴がある。脱落には(1)完全ランダム脱落、(2)ランダム脱落、(3)非ランダム脱落がある。(1)は推計値について統計的に問題ない、(2)は観察可能なデータを用いて対処できる、(3)は、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、対処が極めて難しいとのことである。脱落サンプルを含んだデータは、(1)脱落サンプルを除去する、(2)脱落箇所に数値を補完する（単一値代入法、多重代入法）といった方法がある。他に、(3)利用可能データを最大限生かして分析するものとして、ヘックマンの２段階推定法やパターン混合モデルが紹介されている。

章の後半ではパネル調査のデータセットのつくりかたに触れている。クロスセクションデータにおける変数名や変数の並び、内容が必ずしも複数年のデータベースで整合性がないときもあり、マッチング作業には根気強さと慎重さが要求されるが、この作業を通じてデータの性質もわかるので、かなりの時間と労力をさくべきであると述べられている。

パネルデータは一般に、同一個人の異なる時期のデータが縦に積み重なるような形をしている。ラグの導入などがスムーズにいくように、統計ソフトに、データが時系列データであり、かつ ID も違うデータであることを認識させる機能があるものだと便利である。本書では STATA をつかったプログラム例が紹介されている。時間表示については、分析上、ある年度に行われた何回目の調査か、が重要である場合は、年月日を連続した自然数に置換することが進められている。経済変数のはずれ値については、一般的なルールがあるわけではないが、±４×標準偏差をはずれ値とするという基準が紹介されている。その他、属性ダミーの作成、経済変数のカテゴリー化、経済変数のダイナミックなカテゴリー化などについての有用性などが指摘されている。

（岩澤美帆）

3　パネル調査における統計分析モデル

金子 隆一

1．はじめに

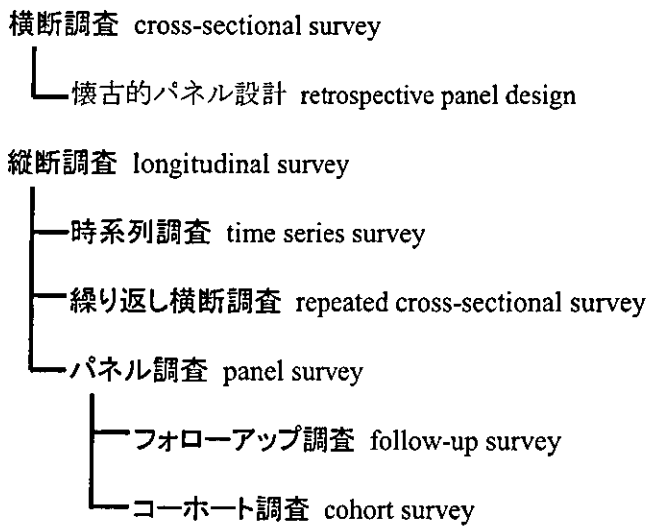
パネル調査(縦断調査)では、同一調査対象を継続的に調査し、その実態や意識の変化を時系列で捉えることによって、人々の行動変化のタイミングや因果関係に対する強力な推論を行うことができるが、その有効性を十分に引き出すためには横断調査とは異なる統計手法を用いなければならない。本稿では、パネル調査分析に必要とされる統計分析法について、その基礎となるモデルを明らかにしながら総括的に論じ、21 世紀縦断調査への応用についての可能性を探る。とりわけ、1990 年代に著しい発展を見せた事象歴分析手法 event history analysis や共分散構造分析手法 covariance analysis(構造方程式モデル structural equation model)は、パネル調査の分析技法の基礎として位置づけられ、本縦断調査の分析においても有効性が期待される。また、ミクロ・シミュレーション技法も今後きわめて有効な分析法となるとみられる。本研究では、基本的な手法を調査し、これらを中心に本調査での有効性、適用法などについての検討の基礎としたい。ただし本稿による本年度の報告においては、基礎事項の概括的把握を中心に記述を行う。

2．パネル調査の統計分析モデル

(1) 調査法と分析デザイン

まず調査法の種別とそこから得られるデータの種類について簡単に整理しておきたい。調査法種別は大きく分けて、横断調査（データ）cross-sectional survey (data)と、縦断調査（データ）longitudinal survey (data)に分けられるが、これらの名称はそれぞれ複数の調査形態の総称と考えた方がよいだろう。まず、横断調査は１時点（あるいは一定の期間）での複数の調査対象に対する調査という点に特徴がある。この測定の同時性、または経時性のなさが後述のように調査事項間の因果関係の特定に対し弱点となっている。一方、縦断調査は、２つ以上の異なる時点において、同一かまたは比較可能な集団に対して、経時的比較を目的に行われる調査である（Menard 1991)。これによれば、同一母集団を想定して反復される横断調査 repeated cross-sectional survey は、縦断調査の一形態と考えられる。同一の課題に対して官庁が定期的に実施する多くの調査はこのタイプである。また、国勢調査などもこれに含まれるだろう。このタイプの調査は、個人単位の調査ではあるが、主に構成比や平均値などで表される母集団の特性のトレンドの観察に用いられる。

図表１　調査の体系



縦断調査の中で、時系列調査は、原則として「単一」の対象、集団について異なる時点で繰り返して実施する調査として分類される。これに対してパネル調査は、「複数」の同一対象（個人）に対して異なる時点で繰り返して実施する調査である[1]。したがって、横断調査における母集団の代表性と、時系列調査における経時性の両面を同時に備えた調査と言える。そのうち、１回の調査で捉えた標本について一定期間後に追加的情報の取得を目的に行われる調査がフォローアップ調査であり、特定の母集団（多くの場合同一年齢層の集団）を定期的、継続的に調査する場合がコーホート調査と呼ばれている。ちなみに、厚労省の行う 21 世紀縦断調査はパネル調査～コーホート調査であり、出生児調査は出生コーホート調査 birth-cohort survey、成年者調査は、年齢コーホート調査 age-cohort survey ということになる。

１回の横断調査においても、調査事項に関する対象者の過去の履歴を調べ、これを時系列データと見なしてパネル調査同様に分析を行う方法があり、懐古的パネル設計などと呼ばれる。しかし、対象者の記憶の不正確さなどから精度に問題が生じやすく、またそれが遠い過去ほど不正確であるなど時間に依存して生ずる点と、過去の意識、態度、理由など心理的な項目を捉えることが困難な点などにおいて、パネル調査に有効性に及ばないとされる。

用語については「縦断 longitudinal」を狭義に用いて、ほぼパネル調査と同義に用いることも提案されている（Baltes and Nesselroade(1979), Wall and Williams(1970)など）が、これら用語については専門家間で必ずしも同意は得られていないと述べている（Menard 1991）。本稿では、複数時点の変数の比較・分析に共通する手法やデザインを指して縦断と

[1] 継続的に保持される対象者一覧表をパネルと呼ぶことからこのように呼ばれる。Paul H. Lazarsfeld の命名とされる。

呼ぶことにし、縦断調査をパネル調査と同義に用いても大過ないであろう。パネル調査の統計分析に際しては、上述の特性を反映して横断調査で一般的に用いられる統計手法（回帰分析に代表される多変量線形モデル）に時系列分析手法を複合して適用することが必要となる。実際、Frees(2004)は、縦断データ分析は時系列分析と回帰分析の結婚であると表現している。

(2) パネル調査の利点ー因果分析

パネル調査の主要な利点として対象（個人）に起こる変化を時系列的に追うことで、変数間の因果関係を統計的に把握できることが挙げられる。要因間の因果関係の特定は、科学的研究の近接的な目的であり、これをもとにして事象のモデル化、科学的理論の構築がなされ、さらには科学的予測 scientific prediction が可能となる。したがって、こうした手続きの方途を与えるパネル調査とその分析は科学的研究において重要な位置づけを持つ。また、政策的観点からは、施策と目標事象・対象との因果関係が明らかにできるから、第一に有効な施策ターゲットの特定が行え、第二に施策を行った効果を事後的に評価することができる。この点でパネル調査は社会科学的な実証分析において中心的な役割を担うものといって過言ではない。

次に因果関係の特定法について簡単に考えよう。一般に一つの変数Xが他の変数Yの変化（変異）の原因であるためには、次の三つの条件が挙げられる。(1)XとYに相関の存在すること（関連性）、(2)XがYに時間的に先行すること（先行性）、(3)相関が見かけ上の関係 spurious relationship ではないこと（竹内 1989, Menard 1991）[2]。見かけ上の関係とは、第３の変数（潜在的独立変数）の介在による相関関係のことであり、要件(3)はXとYが他の変数を介さない直接の関係を持っていること意味する。横断調査では、(1)（相関性）を見いだすことはできる。しかし、(2)（先行性）は一般に正確に捉えることは難しい。懐古的 retrospective に記述された変数を用いて先行関係を特定することもできるが、記憶等に依存する部分は不正確であり、科学的分析としては不十分であることが多い。このように因果関係の要件に時間的要素があることから、純粋な横断調査ではその科学的特定が困難であり、縦断的デザインが必要となる。また、その場合に横断調査とは異なった因果モデルをベースとした統計分析手法が用いられる。以下では、そのパネル調査における因果分析の基礎的な考え方を理解するために、いくつかの基本的モデルを取り上げて見て行くことにする。

(3) 変数変化のモデル

統計的分析の対象として、パネル調査データ(縦断調査 longitudinal data)が横断調査データ

---

[2] これに加えて、異なる対象や時間にわたる普遍性を意味する(4)関連の普遍性または一致性(consistency of association)、理論的な整合性を意味する(5)関連の整合性(coherence of association)も要件とされることがある(竹内 1989)。

(cross sectional data)と最も異なる点は、前者では同一対象を繰り返し調べることによって、関心のある変数の「変化」を明示的に分析の対象とすることができる点であろう。すなわち、変化をモデル上の一つの変数として扱うことができる。

まず、横断調査において二つの変数（X、Y）の因果関係をモデル化する場合を考えよう。Xの値がYの値に対して影響を与えていることを表現すれば、以下のようになる。

$$Y_{i,t} = \beta_{0,t} + \beta_{x,t} X_{i,t} + \varepsilon_{i,t} \qquad (1)$$

ここで、$Y_{i,t}$、$X_{i,t}$は、時刻tにおける個人iの変数値であり、$\beta_{0,t}, \beta_{x,t}$は切片および回帰係数、また$\varepsilon_{i,t}$は$X_{i,t}$と独立に分布する誤差項である。

しかし、横断調査データにおいて、Yの値がXの値にともなって変化していたとしても、それは必ずしも真の「変化」ではなく、時間tにおける個人間の「変異」を変化と見なしていることになる。この変異の中にはもともと個人間に存在する違い（いわゆる個人差）が含まれている。すなわち、変数Yにおける個人差をfとすると、

$$Y_{i,t} = \beta_{0,t} + \beta_{x,t} X_{i,t} + f_i + \varepsilon'_{i,t} \qquad (2)$$

となる（ここではf_iは時間によらないとし、$\sum f_i = 0$とする）[3]。横断調査、すなわち1時点tのみの観察においては、個人差f_iは誤差項$\varepsilon_{i,t}$に含まれ区別することはできないので、もしXが個人差f_iと相関を持つなら、モデル(1)によるXの効果β_xの推定値はバイアス（unobserved heterogeneity bias）を受けることになる[4]。

ところが、これがもしパネル調査によるデータであり、同じ変数に対する調査が以前に（時間t-1とする）行われていたとすると、その2時点間の変化自体をモデル化することができる。すなわち、それぞれの調査時における式(2)を用いて、

$$Y_{i,t} - Y_{i,t-1} = (\beta_{0,t} - \beta_{0,t-1}) + (\beta_{x,t} - \beta_{x,t-1})(X_{i,t} - X_{i,t-1}) + (\varepsilon'_{i,t} - \varepsilon'_{i,t-1}).$$

ここで2時点間の各個人のYの変化を、$\Delta Y_i = Y_{i,t} - Y_{i,t-1}$などと表し、XのYに対する効

---

[3] 式(2)は個人iの効果を切片に含め、$Y_{i,t} = \beta_{i,t} + \beta_{x,t} X_{i,t} + \varepsilon'_{i,t}$と表すこともできる。

[4] 実験などで行われるようにXの値が個人に対して無作為に与えられるような場合には、f_iはXとの独立性が正当化されβ_xは不偏推定量となる。しかし、社会調査においては一般にこれが成り立つことは少ない。その場合には、β_x不偏推定量を得るためには、Xと相関を持つf_i自信か、あるいはこれを表現する観測変数すべて明示的にモデルに入れる必要がある。

果$\beta_{x,t}$が、時間によらない（β_x）と考えると、

$$\Delta Y_i = \Delta\beta_0 + \beta_x \Delta X_i + \Delta\varepsilon_i' \qquad (3)$$

と表され、β_xに対する正しい推定が期待出来る。すなわち、Yの分散のうち個人差に由来する部分を取り除き、変化を正しく評価することができる[5]。こののβ_xはモデル(1)に対する係数β_xと同じものであり(ただし時間によらないと仮定)、式(3)の回帰推定によってモデル(1)が正しく推定できたことになる。

このことは個人差f_iを何らかの個人属性に帰着させたり、あるいは部分的に個人属性によると考えても同じように扱うことができる。すなわち、式(2)におけるf_iの項が、個人属性Uによる効果$\beta_u U_i$に置き換えられるか、あるいは追加されるだけで、2時点間の差を取ると、これらは相殺消去され、結局式(3)に帰結する。つまり、時間変化がないか、あるいは変化の小さい個人属性Uはモデルに取り入れなくともXの効果の推定には影響を与えない。横断的データに対するモデル(1)では、上述のようにf_iを表現しうるようなXと相関を持つ変数をすべて明示的にモデルに入れなければβ_xの推定値はバイアスを持つため、Xの Y に対する因果的関係を統計的に正当化される形で把握することは諦めざるを得ない場合がほとんどである。この点について、縦断データでは、変数の「変化」を明示的に分析の対象とすることができることから、この問題（unobserved heterogeneity、または omitted variables の問題）を回避することができるのである（Frees 2004, Menard 1997 など）。

(4) 変数変化のフィードバック・モデルーstatic-score model

これらは縦断的データの統計的利点を示すための最も簡単なモデルであったが、これらを出発点として、より実用的なモデルに発展させて行く必要がある。パネル調査の強みを示すモデルとして次に紹介するのは、上記の Y の変化が自身の値の大小に依存する場合、言い換えれば自己相関を持つ場合のモデルである。個人の状態 Y が、以前の状態に依存していると考えるのは自然なことである。またその変化の仕方が以前の状態に依存していることもあり得る。たとえば、収入の多かった者は多くを投資に回せるから次期にはより多くの収入が期待される。逆に支出など、たまたま一過的な増加や減少などがあっても継続して調べれば安定化することが期待されるだろう。前者は正のフィードバック、後者は負のフィードバックと考えられる。これらの効果が存在する場合、これらを考慮しないと X の Y に対する因果的関係は正しく推定されない。このモデルは、Y の値が以前の値に依存する形として、以下のように表せる。

$$Y_{i,t} = \beta_0 + \beta_x X_{i,t} + \beta_y Y_{i,t-1} + \varepsilon_{i,t} \qquad (4)$$

---

[5] モデル(3)は、unconditional change-score model、または method of first differences などと呼ばれている。パラメータの標準誤差、検定量等も通常の回帰推定と同様に正しくすいてされる。

ここで回帰係数は時期によらず、また誤差項$\varepsilon_{i,t}$は、$X_{i,t}$および$Y_{i,t-1}$と独立とする。これはある時期のYが、Xだけでなくそれ以前のYに依存して決まること表している。この式は、両辺から$Y_{i,t-1}$を引くことによって、

$$\Delta Y_i = \beta_0 + \beta_x X_{i,t} + (\beta_y - 1) Y_{i,t-1} + \varepsilon_{i,t} \qquad (4')$$

と表すことができる。すなわち、Yの変化がXと、変化前のYの大小に依存して決まることを表している[6]。(4)と(4')では、$X_{i,t}$の$Y_{i,t}$に対する効果、およびYの変化（ΔY_i）に対する効果はともに同じ値β_xであり、（当然であるが）これらのモデルは同等である。β_yは、stability coefficient などと呼ばれる。

(4) 変数変化の一般的モデル

以上に見てきたXのYに対する因果的効果の分析モデルの最も一般的な形式を示せば、

$$Y_{i,t} = \beta_0 + \beta_x X_{i,t} + \sum_{j \in J} \beta_j W_{j,i,t} + \sum_{k \in K} \beta_k Z_{k,i} + \sum_{\tau=1}^{t-1} Y_{i,\tau} + \sum_{\tau=2}^{T} \delta_\tau + f_i + \varepsilon_{i,t} \qquad (5)$$

ここで$W_{j,i,t}$は時間依存属性（変数）、$Z_{k,i}$は時間非依存属性（変数）、J、Kはそれぞれの変数の集合を表し、またδ_tは調査回固有の効果、Tは調査回数を表す（それ以外の記号はこれまでと同じである）。J（時間依存属性の集合）に属す個人属性としては、就業状態・職業、意見・意識など、K（時間非依存属性の集合）に属す個人属性としては、性、出生年、出生地、十分に高い年齢層では身長、学歴などが例として挙げられる。

(5) 変数変化の連続時間モデル

同様の線形回帰モデルの枠組みで、説明要因の効果が目的変量に対して連続的に作用していることを想定したモデルを扱うことができる（Finkel 1995）。すなわち、時間tを連続と考えて、

$$dY_{i,t}/dt = c_0 + c_x X_{i,t} + c_y Y_{i,t} \qquad (6)$$

とすると、これはある時期のYの変化の速度が、同時期のXとYの値に依存して決まることを表しており、c_o, c_xおよびc_yがその依存度を変化率として表すパラメータである。これ

---

6 モデル(4)、および(4')は、conditional change model, static-score model、または lagged-dependent variable model などと呼ばれている。

を微分方程式として期間$(t-\Delta t, t)$について解くと、eを自然対数の底として、

$$\begin{cases} Y_{i,t} = \beta_0 + \beta_x X_{i,t} + \beta_y Y_{i,t-\Delta t} \\ \beta_0 = \dfrac{c_0}{c_y}(e^{c_y \Delta t} - 1)、\ \beta_x = \dfrac{c_x}{c_y}(e^{c_y \Delta t} - 1)、\ \beta_y = e^{c_y \Delta t} \end{cases}$$

となる。これは決定論的モデルであるが、$Y_{i,t}$に対して確率的誤差の項を加えれば回帰方程式と見ることができるから、$\beta_0, \beta_x,$ およびβ_y が推定できる。これらは以下によって式(6)における変化率のパラメータc_o, c_xおよびc_yに変換される（以上 Finkel 1995 による）。

$$c_0 = \beta_0 \frac{\ln \beta_y}{(\beta_y - 1)\Delta t}、\ c_x = \beta_x \frac{\ln \beta_y}{(\beta_y - 1)\Delta t}、\ c_y = \frac{\ln \beta_y}{\Delta t}$$

これら係数は、前回調査時$(t-\Delta t)$から今回調査時（t）の間の$Y_{i,t}$の単位時間あたりの増加率（一定）の要素であり、たとえばc_xはXが1単位大きいときに増加率がどれだけ加算されるかを示す。

(6) 構造方程式モデルの導入－双方向の因果モデル

ここまでXからYへの単純な因果関係のモデルについて見てきたが、現実にはより複雑な状況を扱わなくてはならない場合が多い。因果関係の方向について考えてみよう。まず片方の変数が時間に対して固定的な変数、性別や、出生年、高い年齢における学歴等は、因果的想定においては当然原因、すなわち説明変数にしかならない。しかし、二つの可変な変数を取り上げたとき、たとえば、縦断調査における中心的テーマである女性（妻）の就業状態と出生（子ども数）との関係のように、どちらも他方の原因となり、結果となり得る場合も少なくない。すなわち、社会科学的状況では両方向の因果関係は普通に存在している。また、第3の変数を考えると、これが分析対象とする二つの変数の共通の原因であったり、どちらかの因果を介在したりする場合が考えられる。こうした因果関係は、基本的にそれぞれの変数を同時にしか測定できない横断調査では、検証することが難しい（instrumental variable の利用が必要となる）。

これら複雑な関係を統計的に推定するためには構造方程式モデル structural equation model を用いることになる。構造方程式モデルとは、変数間の因果関係を線形モデルとして表現

したものである[7]。ここではその最も基礎的なモデルとして、２変数の遅延効果のある双方向の因果モデルを見よう。

第１回目の調査におけるそれぞれの値が、第２回目の結果に対して因果関係を持つと想定する場合について考える。X と Y はそれぞれ平均からの偏差で測るものとして、以下のように定式化できる。

$$\begin{cases} Y_{i,t} = \beta_{xy} X_{i,t} + \beta_{yy} Y_{i,t-1} + \varepsilon_{i,t} \\ X_{i,t} = \beta_{xx} X_{i,t} + \beta_{yx} Y_{i,t-1} + \varepsilon_{i,t,x} \end{cases}$$

このモデルにおける X と Y のそれぞれに対する因果の強度を与える係数は、β_{xy} および β_{yx} であり、

$$\begin{cases} \beta_{xy} = \rho_{x_1,y_2} - \rho_{x_1,y_1}\beta_{yy} \\ \beta_{yx} = \rho_{y_1,x_2} - \rho_{x_1,y_1}\beta_{xx} \end{cases}$$

で与えられる。ここで、$\rho_{x_1y_1}$ $\rho_{x_1y_2}$ および $\rho_{y_1x_2}$ は、それぞれ第１回における X と Y の相関係数、第１回 X と第２回 Y の相関係数、および第１回 Y と第２回 X の相関係数である。この結果は、X と Y の時間差をおいた因果強度がそれぞれの相関係数から、自身の因果強度と第１回におけるそれら変数間の相関を乗じた項を差し引いたものとなっていることを示す。すなわち、たとえば X から Y へ因果について調べようとしたとき、$\rho_{x_1y_2}$ の値が大きく一見因果が強く見えたとしても、Y 自身の因果 β_{yy} が強かったり、第１回時点での両者の相関 $\rho_{x_1y_1}$ が大きかったりすれば、真の因果強度 β_{xy} は小さくなることもあり得ることが示される。したがって、二つの調査間の X と Y の二つの相関係数 $\rho_{x_1y_2}$ および $\rho_{y_1x_2}$ の大小を比べることによって因果の強さを判定することはできず、上記モデルによって β_{xy} および β_{yx} を推定して比較する必要がある。場合によっては、X の次期 Y に対する相関の方がその逆

[7] 構造方程式モデルは、潜在変数や同概念を用いた測定方程式モデルと組み合わせることで、より一般的な線型モデルの体系である共分散構造分析モデルへと展開する。それらはパネル調査分析においても重要な役割を果たす。

より強い（$\rho_{x_1y_2} > \rho_{y_1x_2}$）にも関わらず、YのXに対する因果的影響の方が強い（$\beta_{xy} < \beta_{yx}$）場合もあり得ることがわかる。

3．欠損値に対する統計的対処

統計調査、とりわけ回答者自身が記入する形式の調査では、さまざまな理由で回答がなされなかったり、不適切であったりして、一定の標本の変数としてのデータに欠損値が生ずることは避けられない。しかし、上記の議論も含め、一般の統計モデルや理論においてはすべての変数値は揃っていることが前提である。欠損値に偏りがある場合には、これらモデルや理論の前提が整わないことになり、分析の結論に深刻な影響を与えかねない。したがって、欠損値の生じ方のパターンを解析し、偏りの程度などを評価して、統計分析上の適切な対処をする必要がある。偏りが小さいと想定される場合では、多変量解析による回帰係数の推定値に対する影響は少ないとの報告も多く、通常は欠損値を持つ標本を除外することにより分析が行われる。しかしその場合でも、たとえば一つの変数に5％の欠損値が生ずるとすると、95%の標本の情報が利用可能であるが、他の変数について独立にも同様に欠損値が生ずるとすると、わずか10個の変数を用いたモデルに対して、使える標本は60%にまで縮小してしまう（0.95^{10}）。もし、欠損率が10%なら、10変数モデルで使える標本は35%である。こうした大量の標本情報の無駄を防ぐためにも、欠損値に対する適切な方法による補充が必要となる場合がある。以下、欠損値に対する統計的対処について概観しよう。

(1) 欠損のタイプ

統計的な観点から欠損が問題となるのは、欠損に偏りが有る場合、すなわち、その変数あるいは他の変数の値に依存して欠損の生じ方（確率）が異なる場合である。逆に、有る変数の欠損値がその変数の値、または他のいかなる変数の値とも独立に生じている場合は、「完全にランダムな欠損 missing completely at random (MCAR)」と呼ばれ、この欠損を含む標本を除いたデータセットは、もとの標本からの無作為標本となることから、通常の統計手法がそのまま適用できることになる(Allison 2001など)。また、2つの変数XとYを考えたとき、Xをコントロールすると Yの欠損確率が Yに依存しない場合には、「ランダムな欠損 missing at random (MAR)」と呼ばれる。これは Yの欠損が Xの値に依存していても、Xを固定したときに Yの欠損が自身の値にランダムに生じている状況を表している。原則として、通常の多変量解析を提供する際、MCRの条件が満たされているとき（したがって、MCARも含まれる）、欠損値を除いた標本を通常の無作為標本と見なしてより[8]。しかし、

[8] この状況は ignorable と呼ばれる(Allison 2002)。