H. Elizabeth Peters,
"Retrospective versus Panel Data in Analyzing Lifecycle Events,"
The Journal of Human Resources, 23-4, pp.488-513(1988).

本論文では、結婚履歴に関するデータ（就業経験における縦断調査（NLS））について、同一人物による回顧式の設問による回答と、パネルデータによる回答との比較を行っている。

一般的に、回顧式調査のほうが、より完全な結婚履歴をとることができるが、過去のことを思い出すことにともなう誤差が生じやすい。一方、パネル調査各回よって得られた結婚履歴は、記憶による誤差は生じにくいが、次のような理由による誤差が生じやすい。（1）調査間に起こった短期のイベントが観察されない。夫の出生年月を比較することによって、結婚の変化をとらえることも可能だが、夫の出生年月が正確に申告されていない場合、誤った分類をしてしまうおそれがある。さらに（2）パネルデータによって得られる不完全な結婚履歴は、推移確率に関するバイアスのある推定値をもたらすサンプル・セレクションあるいは左センサリングを生じかねない。

実際のデータに基づいて分析したところ、結婚年月については、両者のデータの回答はほぼ一致していた。誤差は系統的なもので、回顧法において思い出すことがだんだん困難になるという事態に関連しているようにみえる。一方、パネルデータでは、調査時点での配偶関係しか訊ねていないので、変化がとらえきれない結婚の存在が明らかになった。

両者のデータによる、結婚年齢の分析においては、推移確率にほとんど違いはなかった。しかし、離婚や再婚のハザード率については、質的には同様の結果が得られたが、パネルデータによる係数の推定値は、回顧式にくらべ、概して不正確である傾向にあった。

（岩澤美帆）

樋口美雄・太田清・新保一成著
「入門パネルデータによる経済分析①～⑤」
『経済セミナー』6月号～11月号（2004年）

本連載は、この分野の初学者を念頭に置き、パネルデータを用いることのメリットや克服しなければならない問題点、さらにはそれを用いるときいにつかわれる分析方法や推定方法などに関する解説がなされている。

(1) パネルデータとは何か（入門パネルデータによる経済分析①）

パネルデータ分析の基本的な考え方は、R.A.Fisher の一連の研究、特に分散分析にさかのぼる。分散分析は、ある外的ショックの与えられたグループ(treatment group)の行動を、ショックの与えられなかったグループ(control group)と比較して、そのショックの影響や効果の大きさを明らかにするための手法である。こうした手法を実際の経済データに応用するには、多数の経済主体について、外的ショックの与えられた前と後における行動変化を調査する必要があり、パネルデータの必要性が古くから認識されていた。

他に先駆けてパネル調査に着手したアメリカでも、それは1960年代に入ってからであった。生活保護制度をいわゆる「負の所得税」制度に変更することによる人々の就業意欲の変化や、税率など給付内容の変更による人々の労働供給行動の変化はいかなるものか。現実のデータにより制度の効果を確認するとの考えから、「負の所得税」に関する社会実験が一部の州でなされた。政策を議論するうえでは、統御実験は不可欠であり、擬似実験から外的ショックに対する人々の行動変化を観察したデータを開発する必要があるとの認識が、当時のアメリカの社会科学者の間でも拡大していた。

こうした背景の中で、パネル調査は始められたのである。例えば、1964年にオハイオ州立大学で National Longitudinal Survey(NLS)、66年にミシガン大学で Panel Study of Income Dynamics(PSID)が始められたのであり、両調査は現在でも続けられている。

パネル調査は同一の個人を追跡調査することによって、複数地点における定点観察の役割を果たしており、人々のダイナミックな行動変化についての情報がわかり、動学分析が可能になる。しかし、成果が出るまでに長い時間と多額の費用を要する。また、ときには調査から得た情報が、全体の母集団を反映しておらず、むしろ歪んだ情報になってしまうという問題もある。

こうした問題点があるが、パネルデータの有効性は強く認識され、アメリカでは60年代、ヨーロッパでは80年代、日本をはじめアジア諸国では90年代になって、パネル調査が実施されるようになった。

(2) パネルデータの利点①（入門パネルデータによる経済分析②）

パネルデータの経済分析上の利点は、大きく分けて 2 つある。第一は、パネルデータで

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$\triangle = Y_t - Y_{t-1}$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$\triangle_t$の相関係数は－１で Y_{t-1}は$\triangle_t$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　パネル調査の統計分析モデルと応用例：イベントヒストリー分析

山口一男 著「イベントヒストリー分析(1)-(15)」『統計』52(9)-53(11) (2002-2003)

鎌田 健司

1．イベントヒストリー分析（event-history analysis）とは

あるイベントを 2 次データにおいて分析しようとする際，横断調査では調査時点におけるイベントが生起したかどうかとそれに付随する属性しか基本的にはわからない。「基本的に」と断わりを入れたのは，横断調査においても結婚や離婚などのイベントを回顧的（retrospective）に発生時点を問うということが一般的になっているからである。しかし，そのような特定のイベントの発生時点がわかったとしても，そのイベントが調査対象者にとってどのようなタイミングにおいて起きたのかを知ることは難しい。縦断調査では同一の対象者を継続的に調査することで，イベントの生起をほぼ時間経過とともに捉えることができ，イベントの発生タイミングについての情報も得ることができる。イベントヒストリー分析は以上のような縦断調査における利点を活かすことのできる有力な分析手法として知られている。イベントヒストリー分析は「確率過程と回帰分析を結びつけたもの」(山口 2002[1])であり，イベントの生起確率とタイミングの両方を考慮した多変量解析である。イベントヒストリー分析は生存分析（survival analysis），ハザードモデル（hazard model）などとも呼ばれる。

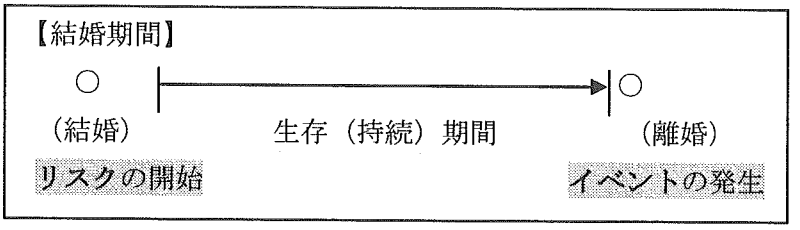
本稿では，イベントヒストリーに関する過去の文献のレビューを中心にイベントヒストリー分析に使用するデータの作成法やパネルデータを用いる場合の留意点などを提示する。

ここで，イベント（event）を「いつ起こった時間（年月や年齢）を特定できる個人の地位や属性，状態の変化」(山口 2002[1]) とし，具体的には結婚，離婚，死亡，就職，離職，転職，移動などのライフイベントのことを指す。例えば離婚をイベントとするとき（図 1），離婚というイベントは結婚というイベントを前提として起こるため，結婚の生起をもって離婚が発生するリスクが開始すると考える。そして結婚時点から離婚が発生するまでの期間を生存期間もしくは持続期間として捉え，イベントが生起するまでの期間を考慮に入れる。イベントヒストリー分析では，イベントが生起するかどうかという情報のほかに，任意に設定された観察期間内においてリスク（risk）開始からイベント発生までの生存（持続）期間（survival time）を考慮にいれて時間経過とイベントの生起するタイミングを分析に取り入れることができる分析手法である（大橋・浜田 1995）。

イベントには，転職，結婚など繰り返し起こるものと，１つのイベントにいくつかの要因が考えられる多義的なイベントが存在し，繰り返しイベントを用いる場合 multiple

duration model といい，多義的イベントを用いる場合 competing risks model という（Allison 1984）。多くの事象は繰り返しイベントであるため，イベントの定義づけを明確にすることでより最適なモデルを構築することができる。

図 1　「離婚」をイベントとしたときのイベント生起の概念図



2．イベントヒストリー分析の利点

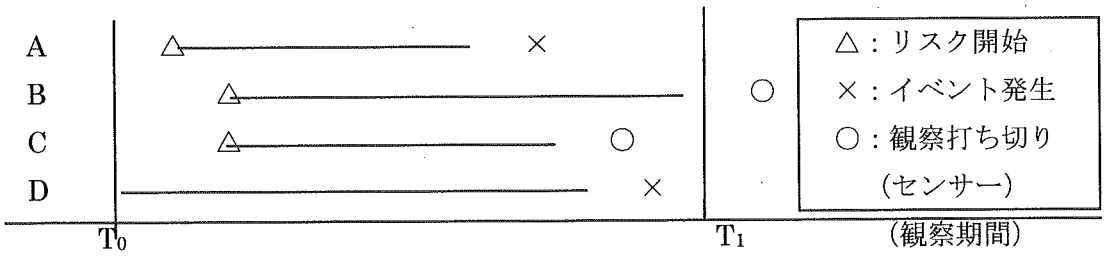
イベントヒストリー分析はイベントが生起する時間経過とタイミングを考慮することができ，イベントをより早いタイミングで生起させる要因の影響力を測定することができるなど，イベントが生起するまでの時間を予測することができることに最大の利点がある。イベントヒストリー分析でイベントのタイミングを考慮できる理由は，イベント数をリスク人口で割った率であるハザード率（hazard rate）を分析に用いる点にある。

山口（2002[1[）では，イベントヒストリー分析の利点を 4 つに分類している。

（1）右センサーされた継続時間の観察値を偏りなく処理できること，
（2）時間とともに変化する予測変数（以下，独立変数）を用いることができること，
（3）他の方法に比べ情報のロスが少ないこと，
（4）時間と予測変数との交互作用効果がある場合にその影響を推定できること 山口 2002[1])

の 4 点について説明している。センサー（censoring）とは，生存（持続）期間中にイベントが生起せず観察が打ち切られることである（図 2)。観察の打ち切りには，2 つの種類があり，それぞれ non-informative なセンサーと informative なセンサーという（大橋・浜田 1995)。センサーされたケースがイベントの生起に関する情報をもっている場合，分析結果にバイアスがかかることがある。センサーがランダムに発生し，イベントの生起に影響がない場合のセンサーを non-informative なセンサーといい，実験法などで予め観察期間を設定する場合をいう。informative なセンサーとは，イベントの生起に影響がある場合を指し，これはさらに右センサー（right-censoring）と左センサー（left-censoring）とに分類される。右センサーは，観察期間にリスクが開始し，観察期間内にイベントが起こらな

図２　イベントの継続時間別の観察パターン：イベントの生起とセンサー



かったケースを示す（図２のB・C)。図２のBは，調査時点までにイベントが生起しないケースであり，図２のDは，イベント生起以外の理由（例えば，ケースの脱落など）によってリスクがなくなり，観察が打ち切られるケースを示す。右センサーされたケースはイベントの生起確率についての情報を持っていると考えられるため，分析モデルに含めることができる。左センサーは観察期間内にリスクの開始が観察されないケースであり（図２のE)，分析モデルに含めることができない。イベントヒストリー分析では，左センサーされたケースが多数存在している場合，分析を適切に応用できないという(山口 2002[1])。このように図２において，Aは観察期間内にリスクが開始し，イベントが起きたケースを示し，B・Cは観察期間内にリスクは開始されるがイベントが起きないケースを示す。Eは分析から除外される。これらを言い替えると，イベントが生起した場合「いつ起きたか」という情報（A）を，生起しなかった場合「いつまで起きなかったか」という情報（B・C）を観測値情報として分析に用いることができるのである。

利点の（2)「時間とともに変化する予測変数を用いることができること」について，山口（2002[1]）は時間とともに変化する予測変数の例として，「個人の血圧，婚姻上の地位や地域の環境汚染度」などの「個人的・社会的リスク要素」をあげている。

利点の（3)「他の方法に比べ情報のロスが少ないこと」，（4)「時間と予測変数との交互作用効果がある場合にその影響を推定できること」については，初婚タイミングの要因研究を例に説明する。イベントヒストリー分析は「確率過程と回帰分析を結びつけたもの」(山口 2002[1]）ということで，重回帰分析（従属変数は初婚年齢）とロジスティック回帰分析（従属変数は初婚経験の有無）との比較で考える。重回帰分析では，調査時点で結婚していないケースを分析に含めることができないため，結婚していないケースの情報をロスすることになる。ロジスティック回帰分析では，初婚のタイミングについての情報をロスすることになり，初婚過程が晩婚による要因によってもたらされたものであるか，早婚の要因によってもたらされたものなのかの違いを区別することができない。また，両分析ともに調査時点の独立変数の値しか使えないため，因果関係成立の独立変数が従属変数に時間的に先行するという要件を満たすことができない。これらの問題はイベントヒストリー分析を用いることで解決できる。まず時間によって変化する独立変数をモデルに含むことが

でき，イベントを経験せずにリスク期間を終えたケースの情報をモデルに反映できる。さらにパネルデータを用いる場合，従属変数に対する独立変数の時間的先行を維持することができ，横断調査を用いる場合よりも独立変数の変数選択の範囲が広がるのである。

3．イベントヒストリー分析に用いるデータ：イベントヒストリーデータ

イベントヒストリー分析に用いるデータ（以下，イベントヒストリーデータ）は，大きく2つの調査データを用いる。1つは，パネル調査に基づくデータを用いる場合と，もう1つは，横断調査において回顧による記録に基づくデータを用いる場合である。イベントヒストリー分析に用いるためのデータとして条件となる基準がある（山口 2002 [2]）。第1に「分析対象となるイベントについてそのリスク開始時間と終了時間とをすべて記録」し，第 2 に「時間で変化する予測変数（独立変数）の収集をしている」ことである。さらに以上の2つの基準は一貫して同じ単位（年，月など）で収集する必要がある。

また，イベントのリスク開始時間と終了時間の収集における留意点として，以下の 3 点について触れている。

(1) 時間を年月で測定する場合，調査対象者の出生年月のデータを手に入れる必要がある

(2) リスクの開始時間が明確でなく，便宜上年齢を基底時間としてハザード率をあらわすイベントについてはイベントの生起時のみ記録し，リスク開始時間は記録しなくてよい

ex.）初婚，初職就業，第1出産

(3) n 番目のイベントの生起が(n+1)番目のイベントのリスク開始時間とみなせる場合は各イベントの生起時のみ記録すればよい

ex.）出産，住居移動（繰り返しのイベントについてもモデル化が可能）（山口 2002[2]）

※ 基底時間：リスク開始時間からの継続時間（いわゆる生存時間）
ex.）離婚の場合，基底時間は初婚期間を示す

イベントが多義的なものである場合，競合するリスク（competing risk）の非独立性の問題が生じる。例えば，死亡をイベントとして扱う場合に，自然死・事故死・病死など死因は様々考えられ，そのような死因の中で事故死を抽出して分析を行いたい場合，その他の理由による死亡は事故死と競合するリスクにあり，それぞれの死因は互いに非独立であるという。この問題を避けるには，事故死以外の死因を右センサーにすることが必要である。これは前提として，「当該イベント（事故死）と右センサーとなるイベント（その他の死因）

が，条件つきで独立」であることを示すことによるものである。ただし，競合するリスクのあるイベントの要因同士に強い相関が見られる場合，このような操作が妥当かどうかは判断がわかれるところである（山口 2002[2]）。

イベントヒストリーデータには 2 種類の形式があり，分析手法によって使用するデータは異なる（表 1）。

表 1　主な分析手法の特徴と対応するイベントヒストリーデータ

分析手法	離散時間ロジット	Cox 回帰	パラメトリック・モデル
データ	Person-period data	Duration data	Duration data
推定法	最尤推定法 maximum likelihood	部分最尤法 partial likelihood	最尤推定法 maximum likelihood
利点	1．通常のロジスティック回帰分析を利用できる 2．時間依存の独立変数を組み込みやすい 3．ハザード率の時間分布を仮定しなくてよい	1．ハザード率の時間分布を仮定しなくてよい 2．3~4つならば時間依存の独立変数を組み込むことが可能 3．連続時間を仮定しているため情報のロスが小	1．イベントが生起する時間を予測するのに適している
欠点	1．パーソン・ピリオド・データの作成に時間がかかる 2．データのサンプル数が多くなる	1．比例ハザード性の仮定を満たす必要がある 2．同じ時間に複数のイベントが多いデータに適さない	1．ハザード率の時間分布に特定のパラメトリックな分布を仮定する必要がある 2．時間依存を考慮できず
備考	連続時間を仮定できる場合，Coxモデルとほぼ同様の結果になる	イベントヒストリー分析で最も一般的な分析法である	指数分布，ワイブル分布，log-logistic分布などの理論分布の仮定が必要

イベントヒストリーデータの 1 つは，パーソン・ピリオド・データ（person-period data）と呼ばれる形式のデータである。

人（person）別・時間区分（period）別のデータということで，パーソン・ピリオド・データと呼ばれる。時間区分のデータが「年」で集計する場合，離散時間モデル（discrete-time）といい，「月」で集計する場合，連続時間モデルという。パーソン・ピリオド・データの利点は，独立変数に時間とともに変化する変数を投入しやすいということがある（山口 2002[3]・[4]）。

山口（2002[4]・[5]）で用いられている例を引用して，パーソン・ピリオド・データの解説を行う。例は，(1)「婚外出産は再婚に不利」,(2)「婚外妊娠だと再婚しやすい」という仮説の検証を目的としたもので，表 2 にはサンプルケースを 6 ケース分示している。ID は同一ケースを表している。観察期間は「離婚もしくは死別後 20 年」をセンサーとしており，月数を基準とした連続時間モデルを用いているため，1 行（レコード）は 1 年を示し，最長同一ケースで 20 行（つまり 20 年分）まで観察される。ST はイベント（再婚）の生起の有

無を示すダミー変数である。再婚が生起すると ST=1 となる。再婚が生起せず右センサーされた場合，ST=0 が 20 行（レコード）分になる。NM は各レコードの経過期間（サンプルデータでは月数）を示している。NM=12 のとき，12 ヶ月分のデータを示す。イベントの生起を表す ST が 1 のときは，NM も必ず 1 となる。イベントが発生したときに NM=3 となっている場合，3 ヶ月連続してイベントが生起するという意味になるからである。従属変数で使用する変数は表 2 の楕円で囲まれたイベント生起を表す ST とその時間経過を表す NM によって示され，残りの DR（リスク開始後何年目か）・CH（離婚死別後各時点まで出産したかどうか）・PG（妊娠状態かどうか）は時間とともに変化する説明変数を示している。例で示された 6 ケースはそれぞれデータ上では以下のような意味を示している。

表 2　パーソン・ピリオド・データの例（山口 2002[4]・[5]）

ID	ST	NM	DR	CH	PG
1	0	12	1	0	0
1	0	12	2	0	0
1	0	12	3	0	0
1	0	7	4	0	0
2	0	6	1	0	0
2	1	1	1	0	0
3	0	12	1	0	0
3	0	12	2	0	0
3	0	9	3	0	0
3	1	1	3	0	0
4	0	11	1	0	0
5	0	8	1	0	0
5	0	4	1	0	1
5	0	6	2	0	1
5	0	6	2	1	0
5	0	12	3	1	0
5	0	7	4	1	0
6	0	2	1	0	0
6	0	10	1	1	0
6	0	12	2	1	0
6	0	12	3	1	0
6	0	4	4	1	0
6	1	1	4	1	0

※ Current Population Survey (1985) 特別調査データ・女性標本のサンプル(山口 2002[4]・[5])

ケース 1 はリスク開始後 3 年 8 ヶ月でイベント（再婚）が生起することなく右センサーされている。ケース 2 はリスク開始 7 ヶ月でイベント（再婚）が生起している。ケース 3 はリスク開始後 2 年 10 ヶ月でイベント（再婚）が生起している。ケース 4 はリスク開始後 12 ヶ月で，イベント（再婚）が生起せずに右センサーとなっている。ケース 5 はリスク開

始後9ヶ月で妊娠（NM=8，PG=1）し，18ヶ月目に出産（NM=6，CH=1）している。ケース6はリスク開始後3ヶ月で出産（NM=2，CH=1）し，3年5ヶ月で再婚している。

パネル調査で得られたデータを用いる場合，例えば調査区間が1年である場合，各レコードは1年分の調査データを示すことになる。

次にイベントヒストリーデータのもう1つの形式は期間データ（duration data）である。これはイベントの履歴が時間を単位とする期間（duration）で表されている形式のデータである。表3は表2のパーソン・ピリオド・データを期間データに作成しなおしたものである。DURはリスク開始（離婚もしくは死別の発生）からイベントの生起（再婚）までの月数を示している。時間とともに変化する説明変数も同様に月数に変換してある。山口（2002[4]・[5]）では，リスクの開始を離婚もしくは死別の発生に設定しているが，例えばイベントを初婚の発生にして，リスクの開始を15歳から49歳という風にリスクの開始を任意に設定することもできる（この操作はパーソン・ピリオド・データでも同様であるが，あまりにリスク期間が長い場合，レコードが膨大な量になる。その点で，期間データであれば，リスク期間を長くとることができる）。

表3　期間データの例（山口 2002[4]・[5]）

ID	ST	DUR	CH	PG
1	0	43	0	0
2	1	7	0	0
3	1	34	0	0
4	0	11	0	0
5	0	43	18	9
6	1	40	3	0

※ Current Population Survey (1985) 特別調査データ・女性標本のサンプルより作成
（山口 2002[4]・[5]）

4．主な分析手法

表1でもふれたが，イベントヒストリー分析は分析対象のイベントの性質・データの時間の単位・収集形式などによって分析手法がいくつかの種類に分かれる。時間の単位で分類すると以下のようになる。

【時間の単位】
- 離散的 ➔ 離散時間ロジットモデル (discrete-time logit model, discrete-time hazard model)
- 連続的 ➔ 等比ハザードモデル（proportional hazard model）
- ➔ Cox モデル（Cox's proportional hazard model）
- ➔ パラメトリックモデル（parametric model）