第 1 章

パネルデータの利点と課題

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、対象に生ずる事象のタイミングや因果関係に対する強力な推論が行える。しかし、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が必要となる。『パネルデータ分析ガイド』は、そうしたパネルデータ特有の分析手法を概説したもので、入門者から本格的な分析研究を目指す者までを対象に実践的なガイドとなることを目指している。わが国では従来パネルデータの蓄積が遅れていたが、近年に至って多くのパネル調査が創設され、分析への関心は高まっている。とりわけ 21 世紀縦断調査は国が行う初の公的パネル調査であり、国民生活の多様な側面を大規模な標本と経時的な調査で捕捉しようとする画期的なものである。本書ではこの 21 世紀縦断調査を中心的な題材としている。ここではまずパネルデータ分析法理解への第一歩として、パネルデータというものの利点と課題について整理をしておきたい。

1.1 調査法と分析デザイン

調査法の種別は大きく分けて、横断調査 cross-sectional survey と、縦断調査 longitudinal survey に分けられる。横断調査は 1 時点における多数の客体に対する調査である。一方、縦断調査は同一または比較可能な客体について、経時的比較を目的に、複数時点で繰り返し実施される調査である (Menard 1991)。縦断調査のうち同一の客体に対して実施する調査はパネル調査と呼ばれる[*1]。調査法分類の一例を図 1 に示した。

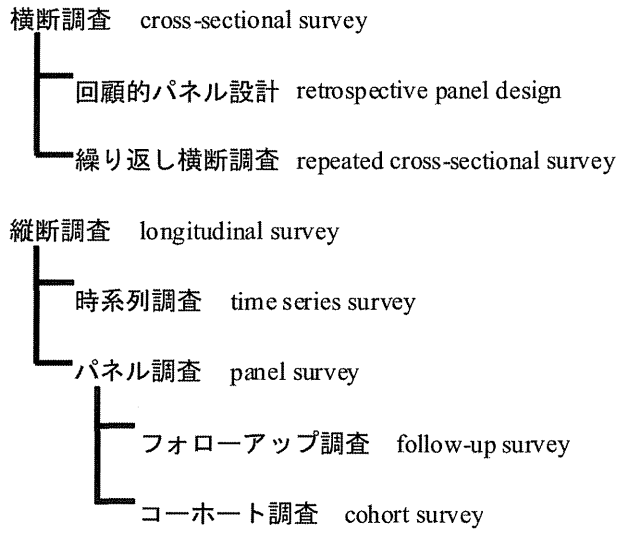
横断調査においても、対象の過去の履歴を調べ、これを時系列データと見なしてパネル調査同様に経時分析を行うことが出来る。これは回顧的パネル設計と呼ばれる。しかし、対象者の記憶に頼るため、遠い過去ほど不正確となるなど時間に依存した誤差が生じやすい点や、過去の意識、意欲といった心理的項目を捉えることが困難な点で、真のパネル調査には及ばない。

繰り返し横断調査とは、同一母集団の変化を捉えるために、異なる時点において異なる標本を抽出して実施するタイプの調査である[*2]。官庁などが同一テーマについて定期的に実施する調査の多

[*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。社会科学における実地調査の開発・発展に寄与した社会学者 Paul H. Lazarsfeld (1901-76) の命名とされる。

[*2] 繰り返し横断調査はその経時性に着目して縦断調査の一形態として分類されることもある (Menard 1991)。

図1 調査の体系



くはこれに属す。

一方の縦断調査の中で、単一の対象を経時的に捉えるケースは時系列調査に分類されるが、社会科学における統計調査では多数の対象を標本として分析することが普通なので、縦断調査の語を狭義に用いて、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）。ただしこれらの語の用法については専門家間でも必ずしも一致を見ない（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呼ぶことにし、縦断調査をパネル調査と同義に用いることとする。

パネル調査の中でも、1回の調査で捉えた標本について、後に追加的情報の取得を目的に行われる調査はフォローアップ調査と呼ばれる。また特定の事象を同時に経験した集団（コーホート）を定期的、継続的に調査する場合はコーホート調査と呼ばれている。ちなみに、厚生労働省の行う21世紀縦断調査はパネル調査の中のコーホート調査に相当し、出生児調査は出生コーホート調査 birth-cohort survey、成年者調査と中高年者調査は、年齢コーホート調査 age-cohort survey ということになる*3。

パネル調査データの統計分析に際しては、上述の特徴を反映して横断調査で用いられる統計手法（回帰分析に代表される多変量一般線形モデル）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。したがって、パネル調査は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。ただし、パネル調査では、調査回を重ねるごとに標本の一部脱落が繰り返され、しだいに標本の代表が損なわれて行くという性質がある。この脱落への対処こそがパネルデータの分析法の最大の課題と言える。これについては後述する。

*3 21世紀出生児縦断調査は2001年1月10日から17日の間及び7月10日から17日の間に出生した子, 21世紀成年者縦断調査は2002年月末時点で20~34歳であった全国の男女及びその配偶者, 中高年者縦断調査は2005年10月末現在で50~59歳である全国の男女をそれぞれ母集団としている。

1.2 パネル調査の利点と因果分析

パネル調査の主要な利点として個々の対象（本稿では個人と呼ぶことにする）に起こる変化を経時的に追うことで、この変化の原因に関する統計的な推論ができることが挙げられる。この要因間の因果関係の特定は、一般に科学的研究の近接的目標であり、これをもとにして事象のモデル化や科学的理論の構築がなされ、ひいては科学的予測 scientific prediction を行うことが可能となる。因果関係の特定は、厳密には科学実験によってのみ可能である。しかし、容易に実験の行えない社会科学の分野では、これに準ずる因果特定の方途を与えるパネル調査とその分析法は重要な位置づけを持つ。とりわけ政策的観点からは、有効な施策の立案・実施は因果モデルによってのみ実現できるものであり、これを目指すための統計調査はパネル型が基礎になるといっても過言ではない。このようにパネル調査は社会科学的な実証分析や科学的根拠に基づいた政策形成において中心的な役割を担うものである。

つぎ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、次の三つの条件を満たす必要がある。(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）[*4]。見かけ上の関係とは、第3の変数（潜在的独立変数）の因果的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）を見い出すことはできる。しかし、(2)（先行性）は一般に正確に捉えることは難しい。回顧的 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、短期的な記憶等に依存する事柄については不正確であり、科学的分析としては不十分となるケースが多い。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（因果関係を前提または想定したモデル）をベースとした統計分析手法が用いられる。では、パネル調査データが横断調査データに比べて因果分析に強いということは、統計モデルから見るとどのように説明できるのであろうか。

1.3 変数変化のモデル

統計的分析の対象として、パネル調査データが横断調査データと最も異なる点は、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。線形回帰モデルによって、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。

[*4] 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)。

$$Y_{i,t} = \beta_{0,t} + \beta_{x,t}X_{i,t} + \varepsilon_{i,t} \tag{1.1}$$

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「差異」を変化と見なしていることになる。この差異の中には、X では説明できない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、個人 i の変数 Y における個人差を f_i とすると、

$$Y_{i,t} = \beta_{0,t} + \beta_{x,t}X_{i,t} + f_i + \varepsilon'_{i,t} \tag{1.2}$$

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）[*5]。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[*6]。

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 1.2 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、

$$\Delta Y_i = \Delta\beta_0 + \beta_x \Delta X_i + \Delta\varepsilon'_i \tag{1.3}$$

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[*7]。 この β_x はモデル 1.1 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 1.3 の回帰推定によってモデル 1.1 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 1.2 における f_i の項が、個人属性 U による

[*5] 式 1.2 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。

[*6] 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 自身か、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。

[*7] モデル 1.3 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推定される。

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、それらは相殺消去され、結局式 1.3 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。このような統計モデルを固定効果モデル（fixed effects model）（解説は第 5 章を参照のこと）という。

横断的データに対するモデル 1.1 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）。

ここで取り上げたモデルは最も単純な形式のものであり、実際の分析では後の章で紹介されるように、より複雑なものを扱わなくてはならない。しかし、パネルデータの利点を活かすための機構についての基本的な考え方は同一であると考えてよい。

1.4 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、回答がなされなかったり、不適切であったりして、データに欠損値が生ずることは避けられない。ところが一般の統計モデルや理論においては、変数値はすべて揃っていることが前提である。もし欠損が特定の値に偏っている場合には、これらモデルや理論の前提が整わないため結論を誤ってしまわないとも限らない。したがって欠損値の生じ方のパターンや偏りの程度を把握して、統計分析上の適切な対処をする必要がある。とりわけパネル調査においては、調査回を重ねるごとに標本には脱落が生じるため、もし脱落が分析対象の変数値に相関して生ずる場合には分析に深刻な影響を与えることになる。したがって、パネル調査分析においては、常に脱落について注意を払っておく必要がある。以下では欠損値に関する課題を簡単に見ておこう。

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、あ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行う際、MAR 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い[*8]。しかし、逆に言うとそうでない場合には、欠損値の統計分析結果に対する影響は無視

[*8] この状況は ignorable と呼ばれる (Allison 2001)。

することが出来ない[*9]。その際には、欠損値の発生パターンに対する統計モデルを特定または想定することによって、一般の統計モデルによる分析法を修正する必要が生ずる。以下ではその対処として欠損値を扱う主な統計手法の種別を挙げておこう。

(1) **欠損値標本の削除 listwise deletion、complete-case analysis**
分析対象となる変数に欠損値を含む標本をすべて分析対象から外す方法であり、一般に最も広く行われている方法となる。様々なタイプの欠損に対して意外に頑健 robust な方法であることが知られている。

(2) **欠損値変数の削除 pairwise deletion、avairable-case analysis**
欠損値を含む変数を分析対象から外す方法であり、具体的には対象とする変数の（平均）分散共分散行列を用いてパラメータの推定を行う。

(3) **ダミー変数法 dummy variable adjustment**
欠損値をカテゴリー変数における一つのカテゴリーと同様に扱い、ダミー変数を立てる方法である。

(4) **代入による方法 imputation**
欠損値に何らかの統計的方法による推定値を代入する方法の総称。具体的な推定方法により様々な方法が考えられる。平均値や多変量回帰モデルの予測値などの単一の値を代入する方法は単一値代入法（single imputation）と呼ばれる。

(5) **最尤推定法 maximum likelihood method**
統計モデルのパラメータの最尤推定の際に、欠損値の発生確率をもとにした尤度を組み込み、欠損値の発生を考慮した推定を行う方法。反復法 iteration method や EM 法 expectation-maximization algorithm などの有効な方法が知られている。

上記の他にも、欠損値の予測そのものが目的ではないものの、多重代入法（multiple imputation）といった方法も知られている。多重代入法では、欠損値にランダムな誤差をもつ値を代入したデータを複数作成し、それらのデータを用いて目的となる統計的分析を行い、最後に複数の分析結果を統合することで、欠損値によるバイアスのないパラメーター（回帰係数）の推定を行うことを目的とする。

1.5 おわりに

パネル調査 (縦断調査) は、実験の困難な人間相手の科学、すなわち医科学や社会科学において、科学的分析の根幹である因果関係の特定に有効な調査デザインであるが、特有の手法の適用を以てはじめてその真価を発揮すると考えられる。『パネルデータ分析ガイド』はまさにそうしたパネルデータ特有の分析手法について実例を付して紹介したものである。もちろん、取り上げていない手法も数多くあるが、パネル調査分析手法一般の基礎について理解を得るように構成されている。本書が 21 世紀縦断調査とともに、わが国のパネル調査の発展と利活用に寄与することを期待する。

[*9] この状況は nonignorable missing と呼ばれる。

参考文献

Allison, Paul D. (2001) *Missing Data*, Sage University Paper series on Quantitative Applications in the Social Sciences, 07-136. Sage, Newbury Park, CA.

Frees, Edward W. (2004) Longitudinal and Panel Data: Analysis and Applications in the Social Sciences, Cambridge Univ. Press.

Menard, Scott W. (1991) *Longitudinal Research*, Sage University Paper series on Quantitative Applications in the Social Sciences, 07-076. Sage, Newbury Park, CA.

竹内 啓 編 (1989)「18 因果分析法」『統計学事典』pp.501、東洋経済新報社.

第 2 章

生存時間分析

2.1 生存時間分析の基本量

パネルデータでは経時観察がされることから、特定のイベント発生までの長さを分析することに利用できるが、このような際に用いられるのが生存時間分析 (survival analysis) である。ここでは生存時間分析に使われる基本的な関数などについて簡単に説明する。[*1]

X をあるイベントが起きるまでの時間を表す確率変数であるとする。生存時間分析では、イベントの生起を死亡にみなして、イベントの起きるまでの時間を生存時間と呼ぶ。このとき、対象がある時間 x を越えて生存する確率は、

$$S(x) = Pr(X > x)$$

で定義されるが、これを生存関数という。生存関数は累積分布関数 $F(x) = Pr(X \leq x)$ と $S(x) = 1 - F(x)$ との関係にある。x が連続で確率密度関数 $f(x)$ が存在するとすれば、

$$S(x) = Pr(X > x) = \int_x^{\infty} f(t)dt$$

となるので、

$$f(x) = -\frac{dS(x)}{dx}$$

が成立する。さらに、

$$h(x) = \lim_{\Delta x \to 0} \frac{Pr(x \leq X < x + \Delta x | X \geq x)}{\Delta x}$$

をハザード関数と呼ぶ。連続な場合には、

$$h(x) = \frac{f(x)}{S(x)} = -\frac{d}{dx} \log S(x)$$

が成立する。

なお、生命表では、$S(x)$ を l_x、$f(x)$ を d_x、$h(x)$ を μ_x と表している。

[*1] 本章の内容については、Klein and Moeschberger (2009)、中澤 (2007) を参考にしている。

2.2 カプラン・マイヤー推定量

Freinreich et al による白血病治療データ (Gehan データと呼ぶ) は、急性白血病にかかり、寛解[*2]の状態にある 42 人の子供について、プラセボを投与する群（対照群）と抗がん剤の 6-MP（6-メルカプトプリン）を投与した群（処置群）に分けて、再発までの時間（月数）を観測したデータである。データでは、各対象者の id（pair)、再発あるいは観察打ち切り（censoring）までの時間 (time)、その時間がイベント（再発）の生起か打ち切りかの別（cens)、そして対照群か処置群かの別（treat）が示されている。まず、データを見てみよう。

Gehan データの読み込みと表示

```
library(survival)
library(MASS)
data(gehan)
gehan
```

出力結果

```
   pair time cens   treat
1     1    1    1 control
2     1   10    1    6-MP
3     2   22    1 control
4     2    7    1    6-MP
5     3    3    1 control
6     3   32    0    6-MP
7     4   12    1 control
8     4   23    1    6-MP
9     5    8    1 control
10    5   22    1    6-MP
11    6   17    1 control
12    6    6    1    6-MP
13    7    2    1 control
14    7   16    1    6-MP
15    8   11    1 control
16    8   34    0    6-MP
17    9    8    1 control
18    9   32    0    6-MP
19   10   12    1 control
20   10   25    0    6-MP
21   11    2    1 control
22   11   11    0    6-MP
23   12    5    1 control
24   12   20    0    6-MP
25   13    4    1 control
26   13   19    0    6-MP
27   14   15    1 control
28   14    6    1    6-MP
29   15    8    1 control
30   15   17    0    6-MP
31   16   23    1 control
32   16   35    0    6-MP
33   17    5    1 control
34   17    6    1    6-MP
35   18   11    1 control
36   18   13    1    6-MP
37   19    4    1 control
38   19    9    0    6-MP
39   20    1    1 control
40   20    6    0    6-MP
41   21    8    1 control
42   21   10    0    6-MP
```

[*2] 症状が一時的に軽くなったり、消えたりした状態のこと

このうち、6-MP を投与したグループについて、時間でソートしたデータを表示してみる。

コードと出力結果

```
gehanT <- subset(gehan, treat == "6-MP")
gehanT[order(gehanT$time),]

   pair time cens treat
12    6    6    1  6-MP
28   14    6    1  6-MP
34   17    6    1  6-MP
40   20    6    0  6-MP
4     2    7    1  6-MP
38   19    9    0  6-MP
2     1   10    1  6-MP
42   21   10    0  6-MP
22   11   11    0  6-MP
36   18   13    1  6-MP
14    7   16    1  6-MP
30   15   17    0  6-MP
26   13   19    0  6-MP
24   12   20    0  6-MP
10    5   22    1  6-MP
8     4   23    1  6-MP
20   10   25    0  6-MP
6     3   32    0  6-MP
18    9   32    0  6-MP
16    8   34    0  6-MP
32   16   35    0  6-MP
```

ここで、イベントが起きた時刻 t_i において、イベントを体験する可能性のある個体数を Y_i、発生したイベントの数を d_i、打ち切りの数を c_i とする。このとき、以下のような表が作成できる。

t_i	Y_i	d_i	c_i
0	21		
6	21	3	1
7	17	1	0
9	16	0	1
10	15	1	1
11	13	0	1
13	12	1	0
16	11	1	0
17	10	0	1
19	9	0	1
20	8	0	1
22	7	1	0
23	6	1	0

そこで、d_i のある時刻だけに改めてインデックス i を振り直し、以下のような生存関数の推定量を考える。

$$\hat{S}(t) = \begin{cases} 1 & (t < t_1) \\ \prod_{i(t_i<t)} \left(1 - \frac{d_i}{Y_i}\right) & (t \geq t_1) \end{cases}$$

例えば、$0 \leq t \leq 6$ では、$\hat{S}(t) = 1$、$6 \leq t \leq 7$ では、$\hat{S}(t) = 1 - \frac{3}{21}$、$7 \leq t \leq 10$ では、$\hat{S}(t) = (1-\frac{3}{21})(1-\frac{1}{17})$ などとなる。これをカプラン・マイヤー推定量 (Kaplan-Meier estimator) と呼ぶ。

カプラン・マイヤー推定量は、R で以下のように推定される。

カプラン・マイヤー推定量

```
gehan.sf <- survfit(Surv(time,cens) ~ treat, data= gehan)
print(gehan.sf)
summary(gehan.sf)
plot(gehan.sf, lty = seq(2),
main = "Kaplan-Meier Plot for Gehan Data")
legend("topright", c("6-MP", "control"), lty=seq(2))
```

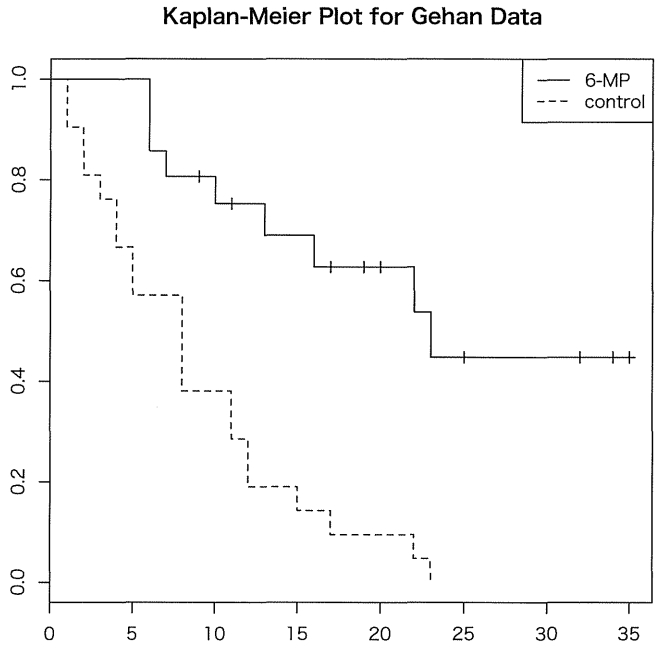
出力結果

```
> print(gehan.sf)
Call: survfit(formula = Surv(time, cens) ~ treat, data = gehan)

              records n.max n.start events median 0.95LCL 0.95UCL
treat=6-MP         21    21      21      9     23      16      NA
treat=control      21    21      21     21      8       4      12
> summary(gehan.sf)
Call: survfit(formula = Surv(time, cens) ~ treat, data = gehan)

                treat=6-MP
 time n.risk n.event survival std.err lower 95% CI upper 95% CI
    6     21       3    0.857  0.0764        0.720        1.000
    7     17       1    0.807  0.0869        0.653        0.996
   10     15       1    0.753  0.0963        0.586        0.968
   13     12       1    0.690  0.1068        0.510        0.935
   16     11       1    0.627  0.1141        0.439        0.896
   22      7       1    0.538  0.1282        0.337        0.858
   23      6       1    0.448  0.1346        0.249        0.807

                treat=control
 time n.risk n.event survival std.err lower 95% CI upper 95% CI
    1     21       2   0.9048  0.0641      0.78754        1.000
    2     19       2   0.8095  0.0857      0.65785        0.996
    3     17       1   0.7619  0.0929      0.59988        0.968
    4     16       2   0.6667  0.1029      0.49268        0.902
    5     14       2   0.5714  0.1080      0.39455        0.828
    8     12       4   0.3810  0.1060      0.22085        0.657
   11      8       2   0.2857  0.0986      0.14529        0.562
   12      6       2   0.1905  0.0857      0.07887        0.460
   15      4       1   0.1429  0.0764      0.05011        0.407
   17      3       1   0.0952  0.0641      0.02549        0.356
   22      2       1   0.0476  0.0465      0.00703        0.322
   23      1       1   0.0000     NaN           NA           NA
```



plot コマンド以下では、出力結果に算出された群別の生存確率をプロットしている。このようにカプラン・マイヤー推定量を算出し、グラフにより表示することで、各集団におけるイベントの発生パターンについて視覚的に把握することができる。

また、両群のイベント発生パターンに統計的に有意な差があるかをログランク検定により検定することができる。ログランク検定では、カイ2乗検定によって、両群の生存関数に有意な差があるのかを検定している。この例では、p値が4.17e-05（=0.0000417）とかなり小さいことから、プラセボを与えられた対照群と抗がん剤を与えられた処置群とでは、白血病の治癒パターンに有意な差があることがわかる。

ログランク検定と出力結果

```
> survdiff(Surv(time,cens) ~ treat, data= gehan)
Call:
survdiff(formula = Surv(time, cens) ~ treat, data = gehan)

              N Observed Expected (O-E)^2/E (O-E)^2/V
treat=6-MP    21        9     19.3      5.46      16.8
treat=control 21       21     10.7      9.77      16.8

 Chisq= 16.8  on 1 degrees of freedom, p= 4.17e-05
```

カプラン・マイヤー推定量を算出する際の注意点として、対象をグループ別に分ける際に、観察期間を通じて変化しない属性を用いる必要があることが挙げられる。カプラン・マイヤー推定量では、ある集団についてイベント発生のリスク開始から終了までのイベント発生パターンを記述している。例えば、各時点の就業状態を使って対象者をグループ別に分けると、リスク期間の途中で就業状態の変化、すなわちグループ間に移動が生じることとなる。すると、ハザード率の分母となる母集団が頻繁に入れ替わることとなり、同一集団の行動としてカプラン・マイヤー推定量を解釈することが難しくなる。

2.3 Cox 比例ハザードモデル

カプラン-マイヤー推定量では、各群におけるイベント発生の経時的なパターンについて把握することが可能である。しかし、各群のイベント発生パターンが複数の要因の影響下にある場合、カプラン-マイヤー推定量では考慮できる要因の数が限られる。また、分析の目的によっては、イベント発生の経時的なパターンよりも、各要因間におけるハザード率の定量的な差異に関心があることも多い。このような場合に用いられる分析手法のひとつが Cox 比例ハザードモデル（Cox' s proportional hazard model）である。Cox 比例ハザードモデルでは、2 つ以上のグループのイベントのハザード率の比較に関心があるとき、両者のハザード関数が比例的な関係にあると仮定し、イベントの生起に影響を与えると考えられる共変量（covariate）によって、ハザード率がどの程度異なるのかを推定する多変量回帰モデルである。

簡単な例を考えてみよう。今、ある年齢層において、男性の死亡に関するハザード関数と女性のそれが比例していると仮定する。女性のハザード関数を基準にとり、$h_0(t)$ と表すことにする。このとき、男性のハザード $h_1(t)$ について、

$$h_1(t) = h_0(t)\exp(\beta)$$

という関係が成立すると考えよう。すると、このパラメータ β を推定することで、両者のハザード関数の差を定量的に評価できることになる。ちなみに、ここでは男性の死亡ハザード $h_1(t)$ は、女性の死亡ハザード $h_0(t)$ に比べて、リスク期間の すべての時点において$\exp(\beta)$ 倍高いと推定されている。$\exp(\beta)$ は基準となる集団と当該集団のハザードの比を表すことから、ハザード比（hazard ratio）といわれる。

次に、これに教育水準の高低を加えたいとする。この場合、女性・高学歴を基準ハザードとし、今度は、性別が男性なら 1、女性なら 0 となる変数 z_1、教育水準が低学歴なら 1、高学歴なら 0 となる変数 z_2 を考えると、

$$h(t|z_1, z_2) = h_0(t)\exp(\beta_1 z_1 + \beta_2 z_2)$$

という関係を仮定して、パラメータ β_1、β_2 を推定することにより評価が行えることになる。例えば、高学歴女性の死亡ハザードを $h_0(t)$ とすると、低学歴女性の死亡ハザードは $h_0(t)\cdot\exp(\beta_2)$ であり、高学歴男性の死亡ハザードは $h_0(t)\cdot\exp(\beta_1)$、低学歴男性の死亡ハザードは $h_0(t)\cdot\exp(\beta_2)\cdot\exp(\beta_2)$ として表される。このようなモデルが Cox 比例ハザードモデルである。より一般的には、

$$h(t|Z) = h_0(t)\exp(\beta' Z)$$

という形となる。

R では coxph という関数を使って比例ハザードモデルのパラメータ推定を行えるようになっている。そこで、Gehan データにおいて、6-MP 処置群を基準ハザード $h_0(t)$ に取り、対照群のハザードが $h_1(t) = h_0(t)\exp(\beta)$ と表されるとした場合のパラメータ推定を行ってみよう。

Cox 比例ハザードモデルと出力結果

```
gehan.ph <- coxph(Surv(time,cens) ~ treat, data= gehan)
summary(gehan.ph)

Call:
coxph(formula = Surv(time, cens) ~ treat, data = gehan)

  n= 42

               coef exp(coef) se(coef)     z Pr(>|z|)
treatcontrol 1.5721    4.8169   0.4124 3.812 0.000138 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

             exp(coef) exp(-coef) lower .95 upper .95
treatcontrol     4.817     0.2076     2.147      10.81

Rsquare= 0.322   (max possible= 0.988 )
Likelihood ratio test= 16.35  on 1 df,   p=5.261e-05
Wald test            = 14.53  on 1 df,   p=0.0001378
Score (logrank) test = 17.25  on 1 df,   p=3.283e-05
```

ここで、exp(coef) で示されているのが $\exp(\beta)$ の値に相当する。対照群のハザード関数は、基準ハザードに比べて約 4.8 倍も高いものとなっていると推定され、両群のハザード率は統計的に有意に異なっている。

このとき、対照群と処置群のハザード関数（時間の経過によるイベントの発生パターン）は同じ形状であり、その水準（あるいは高低）のみが比例的に異なるということを仮定していること（＝比例ハザードの仮定）に注意が必要である。

また、Cox 比例ハザードモデルでは、部分尤度法（partial likelihood method）を用いて、パラメーターの推定を行うが、部分尤度法においては、イベントの生起について連続時間を仮定している。そのため、Cox 比例ハザードモデルでは、パラメーター推定の計算過程において、同時に複数のイベントが発生すること（tied event）を想定していない。しかし、Gehan データをみると明らかなように、実際には同じ月で複数のイベントが発生している。同時イベントの問題は頻繁に起こりうるため、一般的な統計ソフトではエフロン法（Efron method）、ブレスロウ法（Breslow

method)、厳密法 (Exact method)、離散法 (Discrete method) などの近似を用いてパラメーターを推定する (Allison 1995)。R では標準的にエフロン法が用いられているが、オプションによりその他の近似法を選ぶことも可能である。

Allison (1995) によれば、同時発生イベントが少ない場合には、どの近似法を用いても推定結果に大きな違いをもたらさないが、時間の測定単位が粗い (例えば、年単位) こと等によって、サンプルに対して同時発生イベントが多い場合には、用いる近似法によって異なる結果を得る恐れがあるという。エフロン法とブレスロウ法では、計算負荷が小さく、推定にかかる時間が少ないという利点があるが、同時発生イベントが多い場合には係数の値が 0 に近づく形でバイアスがかかることが指摘されている。厳密法や離散法ではこのようなバイアスは生じないが、サンプル数が増えるに従って、計算負荷 (推定にかかる時間) が飛躍的に上昇するという難点がある。ただし、エフロン法はブレスロウ法よりもバイアスが少ない推定結果を得ることから、少なくとも探索的な分析の段階では、エフロン法を用いることが推奨されている。なお、厳密法と離散法とでは、ほぼ同様の結果を得るが、厳密法の方が計算負荷が高い。両者の違いは、同時発生イベントについて、時間を離散的なものとみなし、離散的な時間の中で真に同時にイベントが生起したとみなすか (離散法)、時間は連続的であるが、時間の測定単位が十分に細かくないために同時イベントが発生しているとみなすか (厳密法) の違いであり、その選択はイベント生起に対する本質的な考察に基づくべきであるとされている (Allison 1995)。

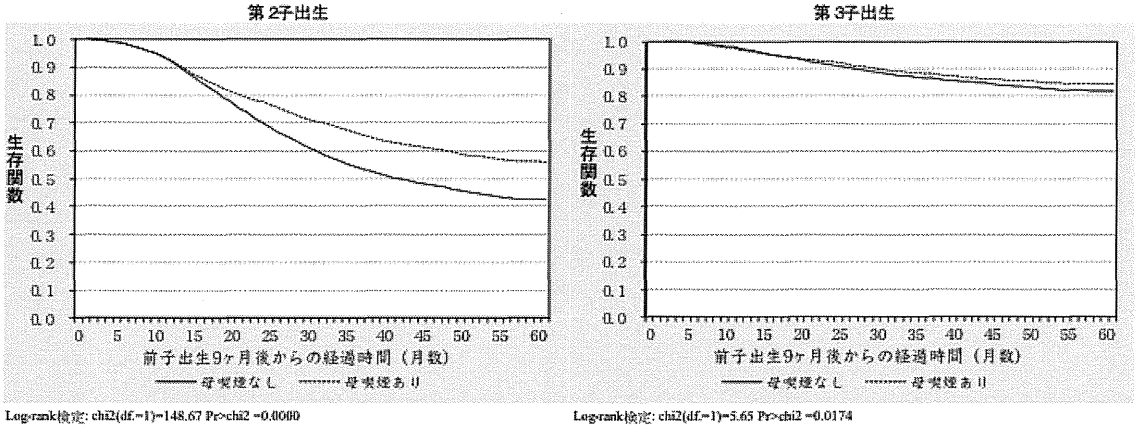
2.4 出生児縦断調査への応用例

コックス回帰モデルを利用した例として、ここでは岩澤・鎌田（2014）による母の喫煙習慣と追加出生との関係を分析した結果を紹介する。

従属変数は、妊娠期間を考慮し、前子出産の9ヶ月後から次子出産までの待ち時間であるが、次子の出生がないまま観察期間が終わってしまった場合や縦断調査の対象者が観察の途中で脱落してしまった場合も、コックス回帰モデルの場合はセンサリングケースとして観察が終わるまでの情報を生かすことができる。検証したい効果は、前子出産後半年後に母親に喫煙習慣があった場合、喫煙習慣がなかった母親に比べ、次子の出生タイミングにマイナスの関係があるかどうかである。

出生タイミングに違いをもたらす様々な共変量を統制したコックス回帰モデルの推定の前に、母の喫煙の有無のみによって出生タイミングがどの程度異なるのかを、カプラン・マイヤー法によって確認しておくことが重要である。下図はカプラン・マイヤー法による次子出生についての生存関数（各経過時点で事象が起きていない割合）である。母の喫煙習慣の有無によって生存関数に差がないという帰無仮説をカイ二乗検定で検証するログランク検定によると、第2子出生については1%未満の水準で統計的に有意な違いがあり、第3子出生については、第2子出生ほど違いは顕著ではないが、5%未満の水準で統計的には差がある、すなわち母に喫煙習慣があった場合は有意に次子出生タイミングが遅いという結果が得られた。

カプラン・マイヤー法による次子出生についての生存関数

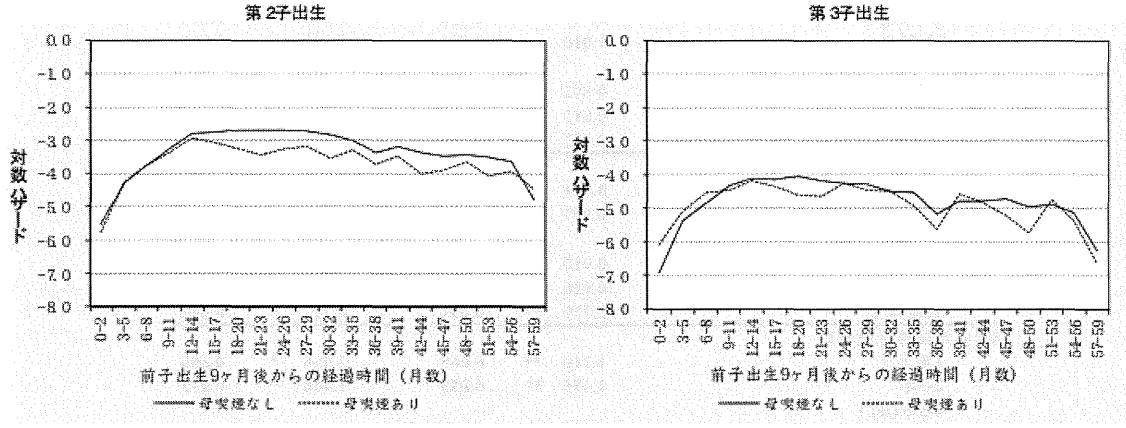


出所：岩澤・鎌田（2014）

さらに、コックス回帰モデルでは、共変量の効果、すなわちハザード比が経過時間に関わらず一定である比例ハザード性を仮定している。母親の喫煙習慣の効果についても比例ハザード性が仮定できるかを調べるために、対数ハザードを図示し、母親の喫煙習慣の有無による差が経過時間に関わらず一定（両水準が平行）かどうかを確認しておくことも重要である。下図に3ヶ月を単位として算出した対数ハザードを示したが、第2子、第3子ともに二種の対数ハザードは概ね平行であることがわかる。対数ハザードがはっきりと交差していたり、差の幅が時間によって大きく変動している場合には、経過時間との交互作用項を含めることでハザード比が時間によって変わることを許

容するモデルを利用することが望ましい。

次子出生についての対数ハザード



出所：岩澤・鎌田（2014）

コックス回帰モデルの推定結果を表に示した。第2子については母親の喫煙習慣の効果は1％未満の水準で有意であり、喫煙習慣のない場合のハザードに比べ74.8％の水準であることがわかる。また第3子については5％水準で差が有意であり、喫煙習慣のない場合に比べ81.9％の水準であった。

次子出生タイミングについてのコックス回帰分析の結果

共変量	第2子 ハザード比		Std. Err.	第3子 ハザード比		Std. Err.
母結婚年齢						
16-22歳	0.907	*	0.038	1.240	*	0.094
23-25歳	1.010		0.026	1.102	+	0.057
26-28歳(ref.)						
29-31歳	0.892	**	0.027	0.918		0.063
32-34歳	0.815	**	0.037	0.676	*	0.087
35歳以上	0.473	**	0.035	0.238	**	0.077
父結婚年齢						
18-22歳	0.924		0.047	0.911		0.087
23-25歳	0.997		0.030	0.946		0.056
26-28歳(ref.)						
29-31歳	0.910	*	0.025	0.931		0.054
32-34歳	0.926	*	0.032	0.837	*	0.065
35歳以上	0.779	**	0.031	0.752	*	0.072
第1回時結婚持続期間						
1年以下	1.719	**	0.065	2.056	*	0.606
2-3年	1.475	**	0.053	1.468	**	0.073
4-5年(ref.)						
6-8年	0.731	**	0.045	0.568	**	0.033
9年以上	0.284	**	0.041	0.284	**	0.036
第1子婚前妊娠	0.845	**	0.025	0.957		0.053
母の学歴						
中学校	0.680	**	0.050	0.962		0.128
高校(ref.)						
専修・専門学校	1.049	+	0.029	1.172	*	0.067
短大・高専	1.148	**	0.031	1.004		0.057
大学・大学院	1.136	**	0.037	1.194	*	0.084
父の学歴						
中学校	0.916	+	0.044	0.884		0.084
高校(ref.)						
専修・専門学校	1.038		0.031	1.028		0.066
短大・高専	1.031		0.058	1.189		0.131
大学・大学院	1.038		0.026	1.076		0.058
母対象児出産1年前の就業状況						
無職・学生(ref.)						
常勤（出産後離職）	1.149	**	0.031	1.212	+	0.123
常勤（育休取得あり）	1.095	*	0.035	1.253	*	0.081
常勤（育休取得なし）	1.114	+	0.071	1.130		0.149
パート・アルバイト	1.083	*	0.032	1.055		0.080
自営業・その他	1.010		0.060	1.204	*	0.109
父の就業状況（対象児出生半年後）						
無職・学生	0.702	**	0.064	0.879		0.174
常勤(ref.)						
パート・アルバイト	0.919		0.078	0.939		0.223
自営業・その他	0.994		0.035	1.329	**	0.085
不詳	1.052		0.109	0.888		0.195
父母の親との同居（対象児出生半年後）						
父母の親と別居(ref.)						
母の親と同居	0.934		0.044	0.892		0.085
父の親と同居	1.122	**	0.033	1.158	*	0.063
子育ての不安の有無（対象児出生半年後）						
子どもを持って負担に思う	0.954	+	0.023	0.792	**	0.039
父母の喫煙状況						
父喫煙（対象児出生半年後）	0.925	**	0.020	0.908	*	0.040
母喫煙（対象児出生半年後）	0.748	**	0.024	0.819	*	0.054
Number of obs	20,203			14,306		
LR chi2	1315.16	**		575.63	**	
Log likelihood	-100494.39			-22569.65		
df.	37			37		
AIC	201062.8			45213.3		
BIC	201355.6			45493.3		

Significance level. 0.1 + 0.05 * 0.01 ** (ref.)はリファレンス・カテゴリ

注：前子出生後 9 ヶ月後をリスク期間開始とし、次子出生年月を事象発生とした。脱落ケースはセンサリングとして対象に含んでいる。同時発生ケースの処理は Breslow 法を用いた。

出所：岩澤・鎌田（2014）

参考文献

Allison, P. D. (1995), "Survival Analysis Using The SAS System: A Practical Guide", SAS Institute Inc.

岩澤美帆・鎌田健司 (2014)「縦断調査を用いた出生力の規定要因分析：父母の喫煙習慣効果を検証するモデル比較」金子隆一編『厚生労働科学研究費補助金「縦断および横断調査によるライフコース事象の経時変化分析と施策への対応に関する研究（H24-政策-一般-004)」平成 25 年度総括研究報告書』.

Klein, J. P. and M. L. Moeschberger(2009), 『生存時間分析』, シュプリンガー・ジャパン株式会社.

中澤港 (2007), 『R による保健医療データ解析演習』, ピアソン・エデュケーション.