合、あるいは世帯所得が低い層において、子どもの育ちの不安感が高かった。親が非正規職かどうかで育児不安・負担感の実態がどのように異なってくるのかについては、今後より詳細な実証分析が求められる。

先行研究では、「育児不安・負担感」を一括りにした分析手法を用いたものもある。その一括りにした分析から導き出された命題として、「就業する母親よりも専業主婦の方が育児不安が高い」というものは代表的である。地方自治体における子育て支援の政策現場では、この命題が多く次世代育成支援対策行動計画に引用され、さらに踏み込んで言えば、この命題が政策現場で一人歩きしているようにも思われる。

しかしながら、育児をめぐる否定的な意識（不安、負担、不満等）はそれぞれ次元の異なるものであり、それらを「育児不安・負担感」と一括りにして実証分析すると、かえって実態が見えにくくなる側面があるとも考えられる。

今後、パネル調査を活用した「育児不安・負担感」の把握を行う際は、細分化された概念であるが、時間不足感・身体的疲労感・精神的疲労感・制度面の不足感（保育施設・医療施設）・経済的負担感・配偶者の育児参加不足感・子育てに関する見解の家族内での不一致・子どもの育ちの不安感（健康面、しつけ・育ち面）というような、個別の概念化を行った上での分析も一つの手法だと考えられる。本稿は、個別の概念化を行った上で、各々の項目について、母親の職業別、世帯所得別、父親の労働時間別、父親の子育て時間別、父親の職業別との関連を見るという新しい分析手法の一つの試みである。

【謝辞】　本論文の作成にあたり、数年にわたり研究支援者として、慶應義塾大学大学院・中村亮介氏には多くの支援を頂いた。ここに記して感謝したい。

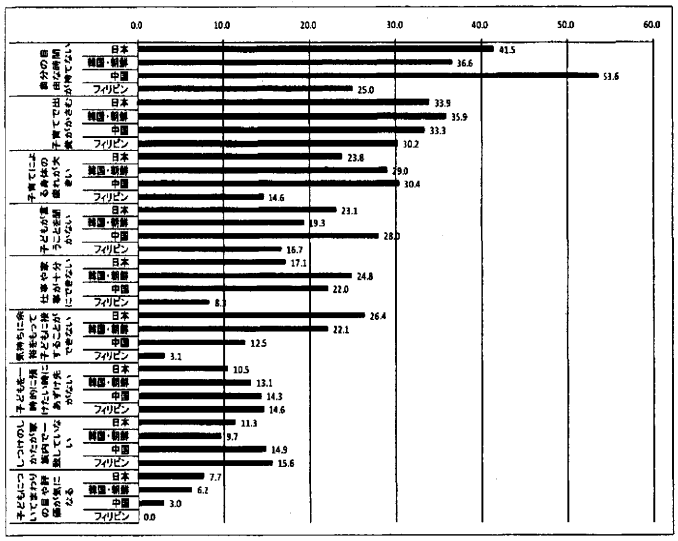
※参考２：母親の国籍別にみた日本の特徴

サンプル数[22]が大きく異なるが、母親の国籍別にみてみると、日本国籍の母親は、「気持ちに余裕をもって子どもに接することができない」「子どもについてまわりの目や評価が気になる」といった、精神的な余裕や周囲の評価について気にする傾向がよみとれる（図表17①②）。

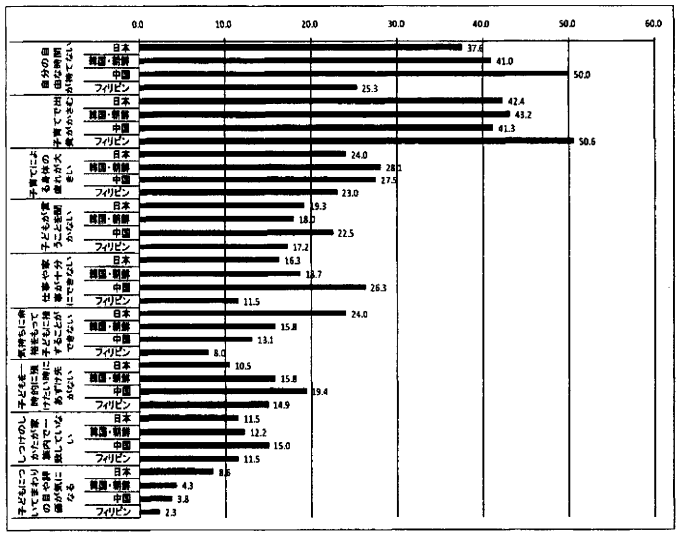
[22] 第5回調査時では、日本国籍の母親が39,324人、韓国・朝鮮籍が145人、中国籍が168人、フィリピン籍が96人である。

図表 20　親の国籍別にみた育児負担・不安感（母親の国籍別）

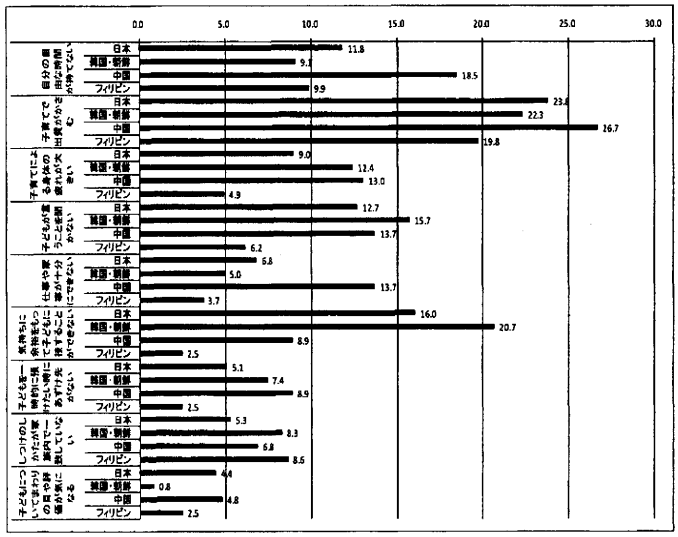
①第 5 回



②第 6 回



③第 7 回



Ⅱ. 個別研究報告

（縦断調査分析方法論の整備
-『パネルデータ分析ガイド』-）

『パネルデータ分析ガイド』への序
（パネルデータの利点と課題）

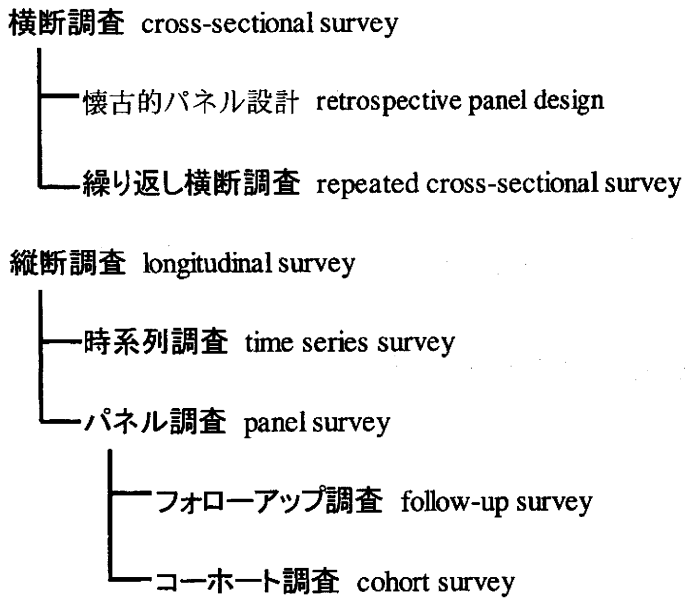
パネル調査(縦断調査)では、同一調査対象を継続的に調査し、その実態や意識の変化を時系列で捉えることによって、対象に生ずる事象のタイミングや因果関係に対する強力な推論が行える。しかし、その有効性を十分に引き出すためには横断調査とは異なる統計手法が必要となる。『パネルデータ分析ガイド』は、そうしたパネルデータ特有の分析手法を概説したもので、入門者から本格的な分析研究を目指す者までを対象に実践的なガイドとなることを目指している。わが国では従来パネルデータの蓄積が遅れていたが、近年に至って多くのパネル調査が創設され、分析への関心は高まっている。とりわけ21世紀縦断調査は国が行う初の公的パネル調査であり、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。

0.1 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）。縦断調査のうち複数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表1に示した。

[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。

図表1　調査の体系



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。これは懐古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ずやすい点や、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

---

[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）。

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。

0.2 パネル調査の利点－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）[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。

---

[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。

[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)。

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とされることが多い。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（因果関係と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。

0.3 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。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、個人 i の変数 Y における個人差を f_i とすると、

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

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）[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

[5] 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。

（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(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、

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

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[7]。 この 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、そ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）。

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『パネルデータ分析ガイド』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。

---

[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。

[7] 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。

0.4 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、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、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）

---

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

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。多変量の場合は、multiple imputation と呼ばれ、複雑となるが、次の最尤推定法と組み合わせた反復法 iteration method である、EM 法 expectation-maximization algorithm などの有効な方法が知られている。

(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。

0.5 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表すと考えられる。『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に寄与することを期待している。

引用文献

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.

パネルデータ分析法ガイド

「パネル調査（縦断調査）に関する統合的分析システムの応用研究」プロジェクト編

目次

はじめに

このパネルデータ分析法ガイドは、厚生労働省統計情報部で実施されている３つの縦断調査の分析を行う上で有用と考えられる、パネルデータに関するいくつかの分析手法の理論とその応用例を示したものである。

第 1 章

固定効果・ランダム効果モデル

本章では、パネルデータ分析の基本的な分析手法である固定効果・ランダム効果モデルについて、統計解析ソフト R による実行例を見ながら統計学的理論を解説し、数値解析例を示すとともに、出生児縦断調査への適用例を紹介する[*1]。

1.1 通常の線形回帰モデル

1.1.1 理論編

いま、N 個の個体を $i = 1, \cdots N$ で表す。変数としては、被説明変数 y_i と、これに対する K 種類の説明変数 $X_i' = \left[\begin{array}{cccc} x_{1i} & x_{2i} & \cdots & x_{Ki} \end{array} \right]$ を考える。このとき、定数項を α、定数項以外の回帰係数を $\beta' = \left[\begin{array}{cccc} \beta_1 & \beta_2 & \cdots & \beta_K \end{array} \right]$、誤差項を u_i として、回帰式は、

$$y_i = \alpha + X_i'\beta + u_i \quad i = 1, \cdots N$$

と書くことができる。

以下、単純化のため、説明変数が一つ、すなわち、$K = 1$ のケースを考える。この場合、上の式は、

$$y_i = \alpha + x_i\beta + u_i \quad i = 1, \cdots N$$

となる。

ここで、残差の平方和が最小になるようにパラメータ α、β を決定するのが最小二乗法 (OLS) である。OLS 推定量は、以下のような仮定の下で、他のいかなる線形推定量よりも分散が小さくなるというよい性質 (BLUE) を持つ (Gauss-Markov の定理)。

1. u の期待値が 0 ($E(u) = 0$)
2. u の分散は均一で、$i \neq j$ について、u_i と u_j は無相関 ($E(uu') = \sigma^2 I$)
3. u は説明変数 x と無相関 ($E(xu') = 0$)

[*1] 本章の内容については、Balatagi (2005)、Croissant and Millo (2008)、北村 (2005)、樋口美雄 [等] (2006) を参考にしている。

α、β の OLS 推定量を $\hat{\alpha}$、$\hat{\beta}$ と書くと、

$$\hat{\beta} = \frac{\sigma_{xy}}{\sigma_x^2} = \frac{S_{xy}}{S_{xx}}$$

$$\hat{\alpha} = \bar{y} - \bar{x}\hat{\beta}$$

となる。ただし、$\bar{x}$, $\bar{y}$ は x,y の平均値、σ_x^2, σ_{xy} は x の分散と、x,y の共分散である。また、

$$S_{xx} = \sum_{i=1}^{N}(x_i - \bar{x})^2$$

$$S_{xy} = \sum_{i=1}^{N}(x_i - \bar{x})(y_i - \bar{y})$$

である。

1.1.2 R による計算（原始的な方法）

1.1.1 節で見た方法をそのまま用いれば、α、β の推定量を求めることが可能である。ここでは、下に示す仮想的なデータである「データセット A」について、OLS 推定量を求める問題を考える。

データセット A は以下のようなデータであり、以下の R による実行例では、"Ydf" と名付けられたデータフレームに格納して取り扱う。

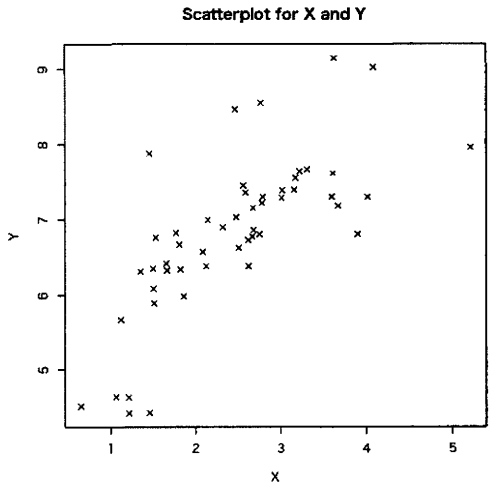
データセット A(データフレーム Ydf) の内容

```
> Ydf
   Ind time        Y        X
1    1    1 6.318302 1.358518
2    1    2 6.765785 1.538198
3    1    3 6.673199 1.812872
4    1    4 7.361044 2.589778
5    1    5 7.305603 2.791172
6    2    1 4.512218 0.652563
7    2    2 4.636834 1.068677
8    2    3 4.630790 1.213461
9    2    4 4.419180 1.215412
10   2    5 4.423203 1.461623
11   3    1 6.633481 2.513192
12   3    2 6.392095 2.623966
13   3    3 6.863346 2.684149
14   3    4 7.303656 4.019452
15   3    5 7.961180 5.209157
16   4    1 5.674991 1.122794
17   4    2 6.089810 1.512576
18   4    3 5.894734 1.518251
19   4    4 5.986789 1.865590
20   4    5 6.809329 3.904122
21   5    1 7.880392 1.463115
22   5    2 8.470850 2.468216
23   5    3 8.556446 2.769034
24   5    4 9.147444 3.630287
25   5    5 9.024920 4.083941
26   6    1 6.348810 1.826399
27   6    2 6.390482 2.129157
28   6    3 6.733463 2.619354
29   6    4 6.777782 2.668968
30   6    5 6.806236 2.758063
31   7    1 6.996998 2.147962
32   7    2 7.032473 2.483352
33   7    3 7.397037 3.164326
34   7    4 7.663331 3.312535
35   7    5 7.613489 3.613653
36   8    1 6.824914 1.773091
37   8    2 7.454234 2.563154
38   8    3 7.390236 3.023997
39   8    4 7.555790 3.179776
40   8    5 7.641978 3.227601
41   9    1 6.358868 1.505719
42   9    2 6.578637 2.088625
43   9    3 7.154565 2.677834
44   9    4 7.222814 2.783386
45   9    5 7.182551 3.674878
46  10    1 6.428335 1.663196
47  10    2 6.335629 1.670820
48  10    3 6.898983 2.324702
49  10    4 7.290997 3.015024
50  10    5 7.305412 3.604515
```

データセットAの変数XとYの関係をプロットすると以下のようになる。

パッケージとプロット

```
plot(Ydf$X, Ydf$Y, type="p", pch=4,
 main = "Scatterplot for X and Y",
 xlab = "X", ylab = "Y")
```



次に、1.1.1節で述べた理論式を直接用いるとの「原始的な方法」によって、α、βの推定量を求めると以下の通りである。

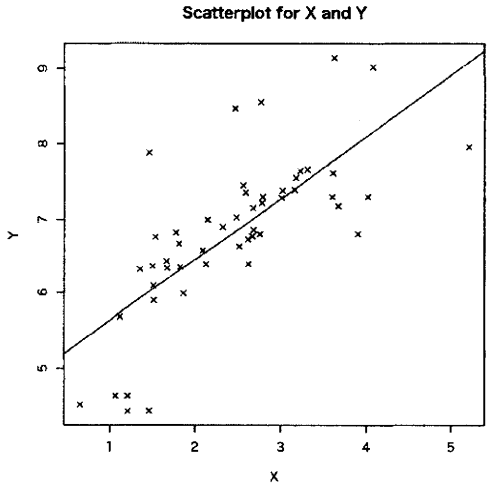
Rによる計算（原始的な方法）

```
V_XY <- cov(cbind(Ydf$X, Ydf$Y))
beta_hat <- V_XY[1,2] / V_XY[1,1]
x_bar <- mean(Ydf$X)
y_bar <- mean(Ydf$Y)
alpha_hat <- y_bar - x_bar * beta_hat
print(c(alpha_hat, beta_hat))

plot(Ydf$X, Ydf$Y, type="p", pch=4,
    main = "Scatterplot for X and Y",
    xlab = "X", ylab = "Y")
abline(alpha_hat, beta_hat)
```

出力結果

```
> print(c(alpha_hat, beta_hat))
[1] 4.807486 0.821806
```



1.1.3 R による計算（関数 lm を利用する方法）

R には線形回帰モデルを推定するための関数 lm が用意されている。これを利用すれば、α、β の推定量のみならず、線形回帰モデルに関する様々な推定量を得ることが可能である。

lm の中には、モデル式といわれる形式で回帰式を記述する。この場合、Ydf というデータフレームの Y を X で説明するという意味になる。説明変数が二つ以上あるときは"+"で結ぶ。結果のサマリーは summary 関数で得られる。

R による計算（関数 lm を利用する方法）

```
lm.ols <- lm(Y ~ X, data = Ydf)
summary(lm.ols)
plot(Ydf$X, Ydf$Y, type="p", pch=4,
    main = "Scatterplot for X and Y",
    xlab = "X", ylab = "Y")
abline(lm.ols)
```

出力結果

```
Call:
lm(formula = Y ~ X, data = Ydf)

Residuals:
     Min       1Q   Median       3Q      Max
-1.58545 -0.26072  0.04754  0.29899  1.87051

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   4.8075     0.2888   16.65  < 2e-16 ***
X             0.8218     0.1100    7.47 1.40e-09 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.7288 on 48 degrees of freedom
Multiple R-squared: 0.5376,Adjusted R-squared: 0.5279
F-statistic:  55.8 on 1 and 48 DF,  p-value: 1.405e-09
```