このエージェントモデルと、縦断調査結果のデータセットを具体的に結びつけるデータクラスを図２に示した。図１における個人（Person)（あるいはその継承クラスの「再生産する個人」PersonReproduce など）は、実はほとんどの属性を各調査回に対応する調査票イメージのデータセットの調査レコード(Record)として保有している。調査レコード(Record)は、たとえば、最終学歴、職業などの多数の調査項目（Item)の集合によって構成されている。さらに、各調査項目（Item)は、その中身としてカテゴリ(Cate）によって構成されており、各カテゴリ(Cate）は基本的に名前と値を持つ。これらは調査結果データを個人別に格納するが、調査レコード（Record）のメソッドを介して、調査結果データセットのファイルから実際のデータを読み込むことになる。これらのメソッドによって、縦断調査結果データとシミュレーションの標本モデルがユーザの見かけ上単純な操作によって、直接に結び付けられる。あとは、個人の振る舞いに関する加齢モジュールに結婚、出生、あるいは就業などの行動モデルを記述することによってライフコース事象が属性や環境に依存しながら発生する様子をシミュレートすることができる。

エージェント型シミュレーションモデルの優れた点は、各種のライフコース事象をそのエージェントが置かれた環境やエージェントどうしの相互作用に依存するモデルを構成し、検証することができる点である。またその依存関係についても境界値による事象の制御など、非線形の関係を記述することができる。これらは通常の統計モデルでは、ほぼ不可能である。実際の個人のライフコース選択においては、環境からの影響やオールオアナッシングの判断などが重要な役割を果たしている可能性があることから、こうした機構を持つモデルについて検証することは、少子化などの現象のメカニズム解明に対して大きな貢献が期待されるところである。また、こうした非線形現象は複雑系現象として知られるが、人間行動が複雑系現象であるとの指摘がなされており、これに対して、21 世紀縦断調査という優れた現実の事象データを用いることができることから、こうした分野の発展にも寄与することが大いに期待できる。

以下には、図２に示した調査結果データのクラスの実装モデル（C++）を示す。

```
//===[ クラス定義 : Cate ]=================================================
class Cate {  // category

  public:

    // ｺﾝｽﾄﾗｸﾀ
    Cate(void);
    Cate(char* str);
    Cate(char* name, int cd, double val);

    // ｺﾋﾟｰｺﾝｽﾄﾗｸﾀ
    Cate(Cate &);

    // ﾃﾞｽﾄﾗｸﾀ
    virtual ~Cate(void);

    // 演算子( = )
    Cate &operator=(Cate &);

    // ﾒｯｾｰｼﾞ操作
```

```
    char   *Name(void);           // 名前取得
    char   *Name(char *);         //  〃  設定
    double  Value(void);          // 値の取得
    double  Value(double);        //  〃  設定

    int     Ncode(void);          // 対応するコードの数を取得
    int     Code(int i);          // i 番目のコードを取得 i の有効性ﾁｪｯｸ後)
                                  // (i=0, 1..., ncode-1, i が範囲外なら-1 を返す)
    int     Code(int i, int c);   // i 番目のコードに c を設定(i の有効性ﾁｪｯｸ後)
                                  // (i=0, 1..., ncode-1, c を返すが、i が範囲外なら-1 を返す)

    void show(void);              // ﾒﾝﾊﾞｰの表示

  private:

    char   *getName(dscr);
    double  getValue(dscr);
    int     getNcode(dscr);
    int    *getCode(dscr);

  private:

    char       *name;          // category name
    double      value;         // category value
    int         ncode;         // # of assigned code to the category
    int        *code;          // assigned code list
};
//===[ クラス定義 : Item ]=================================================
class Item {  // item (=variable)

  public:

    // ｺﾝｽﾄﾗｸﾀ
    Item(char *str);

    // ﾃﾞｽﾄﾗｸﾀ
    virtual ~Item(void);

    // ﾒｯｾｰｼﾞ操作
    char* Name(void);             // 名前の取得
    char* Name(char *n);          //  〃  設定
    int   Loc(void);              // 位置の取得
    int   Wid(void);              // 幅の取得
    int   Type(void);             // ﾀｲﾌﾟの取得

    int    Ncate(void);           // ｶﾃｺﾞﾘ数の取得
    char*  CateName(int c);       // c 番目のｶﾃｺﾞﾘ名の取得
    double CateValue(int c);      // c 番目のｶﾃｺﾞﾘ値の取得
    int    UnknownCate(int* uk);  // 不詳ｺｰﾄﾞを int 配列で取得

    void show(int mode=1);        // mode...Cate を表示するか否か

  private:

    int location(char*); // 位置情報文字列から位置情報設定
    int category(char*); // ｶﾃｺﾞﾘ情報文字列からｶﾃｺﾞﾘ情報設定

  private:

    char*       name;          // 名称
    char*       locstr;        // 位置情報文字列(情報読取り関数が完成したら削除)
    char*       catestr;       // ｶﾃｺﾞﾘ情報文字列(          〃          )

    int         loc;           // location
    int         wid;           // width
    int         type;          // type

    unsigned    ncate;         // # of categories
    Cate**      cate;          // category list

    int         nuk;           // # of unknown categories
    int*        uk;            // unknown category list
```

```
};
//===[ クラス定義 : Record ]==========================================
class Record {  // record

  public:

    // ｺﾝｽﾄﾗｸﾀ
    Record(const char *rdf);          // rdf からすべて読込む
    Record(const char *rdf, char** list);
                                      // rdf から list にある Item を選択して読込む

    // ﾃﾞｽﾄﾗｸﾀ
    virtual ~Record(void);

    // ﾒｯｾｰｼﾞ操作
    char* Name(void);                 // 名前取得
    char* Name(char *n);              //  〃 設定
    char* DataFile(void);             // ファイル名取得
    char* DataFile(char *n);          //     〃     設定
    int   Nitem(void);                // Item の数を取得
    Item* ItemAd(int i);              // i 番目の Item を取得(ﾎﾟｲﾝﾀ)
    char* ItemName(int i);            // i 番目の Item の名前取得
    int   ItemLoc(int i);             // i 番目の Item の位置取得
    int   ItemWid(int i);             // i 番目の Item の幅取得
    int   ItemType(int i);            // i 番目の Item のﾀｲﾌﾟ取得

    int   ItemNmOnlist(int i);        // 呼び出し側から与えられたﾘｽﾄ上での、
                                      // item[i]の位置を取得
    Item* ItemOnList(int j);          // 呼び出し側から与えられたﾘｽﾄ上 j 番目の
                                      // item を取得

    // ﾒﾝﾊﾞｰ表示
    void show(int mode=1);            // mode... Cate を表示するか否か

  private:

    int  readRdf(const char *rdf, const char** list);
                                      // ﾚｺｰﾄﾞ記述ﾌｧｲﾙの読込み
                                      // list にある Item のみ選択して読み込む
                                      // list=0 ならすべて読み込む
    int  isItemIn(const char* istr, const char** list);
                                      // Item 記述の中の Item 名が list にあるかどうか check
                                      // 有れば位置を、無ければ-1 を返す
                                      // : list の最後の要素は空文字列

  private:

    char       *name;                 // record name
    char       *datafile;             // data file name(full path)
    int        nitem;                 // # of variables
    Item       *item[MAXITEM];        // Item list
    int        itnm[MAXITEM];         // 各 item が、呼び出し側から与えられたﾘｽﾄ上の
                                      // 何番目(0, 1...)を記録
};
```

まとめ

本研究では、21 世紀縦断調査データを用いたライフコース事象のマイクロシミュレーション分析を行うための基礎的システムの検討・設計を行った。マイクロシミュレーション分析は、パネル調査との親和性が強く、既存の統計分析ではできない非線形現象としての事象メカニズムの分析や、脱落等の評価が行えるため、統計分析との併用によって縦断調査データの活用範囲を広げるとともに、提供する情報の信頼性向上に資することが期待できる。次年度はここで設計したシステムを実装し、実際のシミュレーション分析を行う。

2　パネル調査の統計分析モデル

（1） パネル調査の統計分析モデル：共分散構造分析：（McArdle J. John and Fumiaki Hamagami, 2006）レビュー

鎌田 健司

1．はじめに

本稿では，パネル調査の統計分析手法として共分散構造分析を取り上げる。共分散構造分析は構造方程式モデル（Structural Equation Modeling：SEM）とも呼ばれ，平均・分散・共分散を用いて変数間の関連を測定するモデルである。この分析手法は，重回帰分析（パス解析）と因子分析の性質を複合的に用いることができ，主に心理学などで用いられる手法である。しかし，人口学分野においても近年，価値観変動に関する分析も散見されるようになり，利用可能で有用な分析手法であると考えられる。

本報告は，2006年6月29日から7月2日まで東京大学と慶應大学において開催された，東京大学総括プロジェクト機構 ジェロントロジー寄付研究部門慶應義塾大学経済学研究科・商学研究科連携21世紀COEプログラム「市場の質に関する理論形成とパネル実証分析」共同開催ワークショップ（Longitudinal Research Institute Workshop “Structural Equation Modeling in Longitudinal Research”, John J. McArdle and Fumiaki Hamagami）のレジュメのレビューを中心に説明する。

概要は以下の通りである。はじめに，共分散構造分析の簡単な説明として，構造方程式と測定方程式，観測変数と潜在変数，構造変数と誤差変数，外生変数と内生変数など特徴的な変数構成についてまとめる。

その上で，パネルデータを用いた場合のモデルをいくつかレビューする。最単純なモデルとして，2時点における構造方程式モデル（two-occasion longitudinal data with SEM）を取り上げる。

次に欠損値を含んだ不完全なパネルデータ（unbalanced panel data）を用いる場合の対処法やモデリングについてまとめる。完全なパネルデータ（balanced panel data）を用いた場合の推定値と不完全なパネルデータを用いた場合の推定値，欠損値を含まないデータのみの推定値の差を補完するための手法などについてまとめる。

さらにカテゴリカル変数を用いたモデルについてまとめる。社会科学の分野においてカテゴリカル変数は最も一般的である。しかし，平均，分散，共分散を用いる共分散構造分析においてはこのような変数は使い勝手が悪かった。ここでは，2値変数を連続変数として変換するテトラコリック相関（tetrachoric correlation）を用いた場合のモデルについてまとめる。

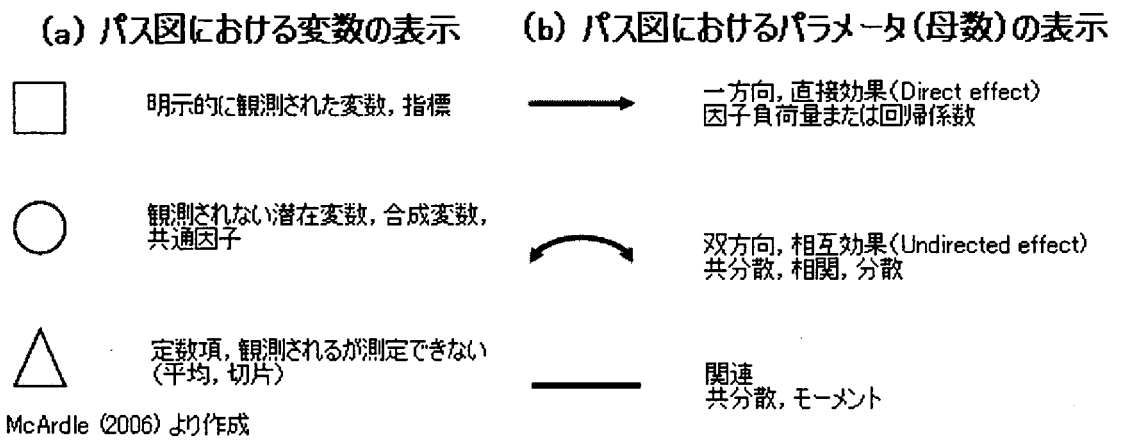
2．共分散構造分析の基礎概念

共分散構造分析は，平均・分散・共分散を用いて変数間の関連を測定する分析手法である。共分散構造だけではなく平均構造も分析できることから，共分散構造分析という名称よりも構造方程式モデル（Structural Equation Modeling：SEM）と呼ばれる方が一般的である。このとき「構造方程式」とは，「構成概念間の因果関係を記述する方程式」（竹内・豊田 1992）であり，複数の因果関係を同時に表現することに特徴がある。構造方程式によって示された変数群をそれぞれ測定可能な形に変換した式を測定方程式と呼ぶ。

共分散構造分析は，重回帰分析（パス解析）と因子分析の性質を拡張したモデルということが特徴であり，因子分析において主に使用される潜在変数（latent variable）をモデルに組み込むことができる。潜在変数とは，実際に観測される観測変数（observed variable）と対をなす変数であり，観測変数群に共通する因子（共通因子）として想定されるほか，誤差変数もこの種類の変数である。構造方程式内で推定される変数を構造変数（structural variable）といい，その誤差項を示す変数を誤差変数（error variable）という。また構造方程式内で他変数から因果関係を指定される場合，その変数を内生変数（endogenous variable）といい，そうでない変数を外生変数（exogenous variable）という。

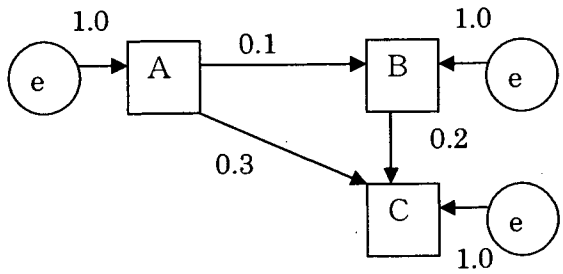
図 1 は構造方程式をグラフ化（パス図）した場合の表示形式を示したものである。四角は明示的に観測された変数を示し，丸は観測されない潜在変数，合成変数，共通因子を示し，三角は観測されるが測定できない変数として定数項（変数，切片）を示す。パラメータの表示については，一方向の矢印は直接効果（direct effect）を示し，推定方法によって回帰係数や潜在変数を想定した場合の因子負荷量を表す。双方向の矢印は相互効果（Undirected effect）を示し，推定方法によって共分散，相関，分散を表す。矢印無しの棒線は変数間の関連を示し，推定方法によって共分散，モーメントを表す。

図 1　構造方程式をグラフ化（パス図）した場合の表示形式



共分散構造分析における変数間の効果は，直接効果，間接効果（indirect effect），総合効果（total effect）に分類される。図2のような構造方程式があるとき，観測変数Aから観測変数Cへの直接効果は偏回帰係数である0.3である。観測変数Aから観測変数Cへの間接効果（観測変数Bを経由した場合）は，観測変数Aから観測変数Bへの直接効果（偏回帰係数）と観測変数Bから観測変数Cへの直接効果（偏回帰係数）を掛け算した値0.02となる。観測変数Aから観測変数Cへの総合効果は，直接効果と間接効果を足した値となるため，0.32というように計算することができる。

図2　直接効果，間接効果，総合効果の計算例



A→Cの直接効果：0.3
A→Cの間接効果（Bを経由）：0.02
0.1 × 0.2 = 0.02
A→Cの総合効果：0.32
0.3 + 0.02 = 0.32

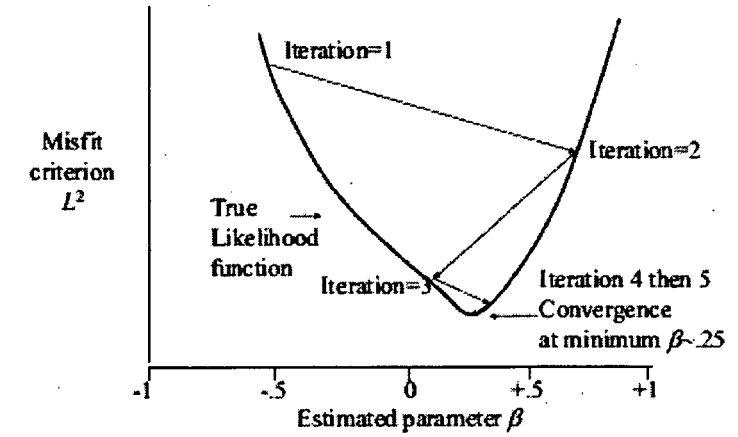
3．パネルデータを用いた共分散構造モデル

3－1．共分散構造モデルの分析過程
(Section B: Structural Equation Models for Change over Two-Occasions)

共分散構造分析を行う場合，推定方法は一般的に回帰分析やANOVAを用いるが，共分散構造分析用の統計ソフトがいくつかある。Joreskog & SorbomのLISREL，Neale at MCVのMx，Muthen & MuthenのMplusなどである。一般的に普及している統計ソフトにおいても，例えばSPSSのAMOS，SASのCALIS，BMDPのEQSがあり，さまざまな環境において共分散構造分析が可能である。

McArdle(2006)のSecsion Bでは，共分散構造モデルの分析過程を以下の4つのステップに整理している。第1段階は特定化（Specification）である。モデルを構築するための仮説が予め特定化されている必要がある。第2段階は期待（Expectation）である。仮説から導き出される変数間の構造方程式・測定方程式の形式で定義し，統計ソフトに入力するという過程である。

図 3　反復推定の概念図



McArdle (2006) より転載

第 3 段階は推定（Estimation）である。定義された構造方程式を統計ソフトにて係数や標準誤差などの統計量を推定する。ＳＥＭでは一般的な回帰分析とは異なり，反復推定（iterative solution）を行うことによって推定値を導き出す。反復解とは，母数を適当な値からスタートさせて，モデルの適合度によって評価するための値である。適合度関数（the fitting function）によって与えられた値から反復計算を行い，モデルの適合度が高いモデルを探る。モデルの適合度が最も高くなったところで計算が終わり，値が確定する（図 3）。

第 4 段階は再検討（Review）である。推定されたモデルをその他のモデルと比較し，モデルの説明力などの精度を高めるための試行錯誤を行う。構造方程式モデルにおいては，期待値 (expected statistics) と観測された統計量とを比較し，モデルの適合度を評価する。つまり，残差の分散を直接的に最小化させるのではなく，モデルによって推定された統計量（＝期待値）とデータから得られる統計量の差を最小化させることでモデルの適合度を高めるのである。このようなモデルの評価は尤度（likelihood）を算出することによって行われる。各個人に対する対数尤度 (log likelihood) は図 3 の Misfit criterion (L^2)に相当し，$L^2=N^*\{[m-\mu],[C-\Sigma]\}$ で示される（m：観測された平均，μ：平均の期待値，C：観測された共分散，Σ：共分散の期待値）。

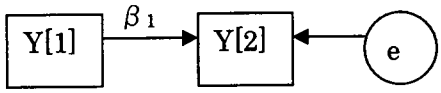
3－2．2時点における構造方程式モデル（two-occasion longitudinal data with SEM）

ここでは 2 時点における構造方程式モデルを取り上げる。このモデルは，従属変数の時間経過による変化を推定するものである。従属変数が繰り返し測定したデータ（repeated measured）かどうかによって利用できるモデルも異なる。

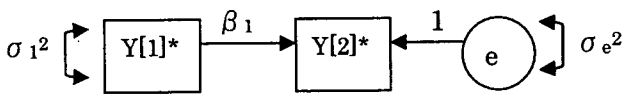
(a) 繰り返し測定するデータを用いた自己回帰モデル
(A typical auto-regression path model for two repeated measures)

はじめに繰り返し測定するデータの最単純モデルを想定して，基本的な事柄を説明する。1 時点におけるイベントの測定をY[1]，時間経過を経て測定されたイベントをY[2]とし，Y[2]を従属変数としY[1]で自己回帰（auto-regression）モデルである。以下の方程式はサンプル 1 からNまでの線形モデルとパス図である。β_0は切片であり，Y[1]=0 のときの予測値となる。はY[1]が 1 単位増加したときのY[2]の変化量である。e は残差である。

$$Y[2]_n = \beta_0 + \beta_1 Y[1]_n + e_n$$

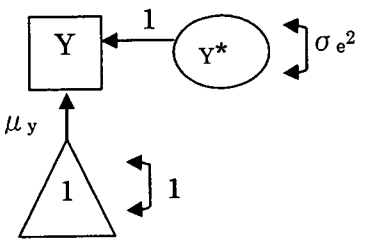


次に共分散を用いた自己回帰モデルを想定すると以下のようになる。アスタリスク（*）は平均周辺の偏差を示す。

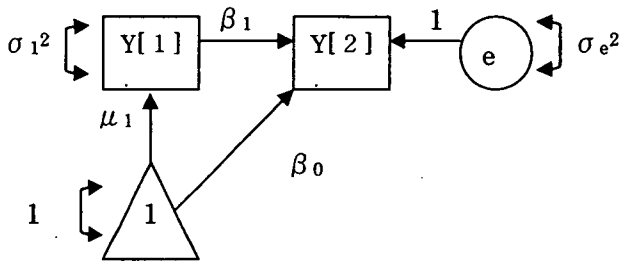


ここで，Y に対する平均と分散は以下のように定義される。μ_0はグループ内の定数項を示し，Y_n*は各サンプルの平均周辺の偏差（$Y_n - \mu$）を示す。

$$Y_n = \mu_y + Y_n^*$$
$$\Sigma \{Y_n^{*2}\} / (N\text{-}1) = \Sigma \{Y^* Y^*\} = \sigma^2$$



最後に，最単純な線形自己回帰モデルに平均と分散を考慮したモデルは以下のようになる。定数項から従属変数までの係数は平均を示し，独立変数までの係数は切片を示す。



自己回帰モデルによって得られた推定値の解釈は以下のようになる（図4）。

図4　自己回帰モデルによる推定値の解釈

自己回帰モデルによる推定値の解釈

$$Y[2]_n = \beta_0 + \beta_1 Y[1]_n + e_n$$

1. **残差変化(residual change)**

$$(Y[2]_n - \beta_1 Y[1]_n) = \beta_0 + e_n$$

2. **直接変化(direct change)**

$$(Y[2]_n - Y[1]_n) = \beta_0 + \beta_1 Y[1]_n + e_n - Y[1]_n$$
$$= \beta_0 + (\beta_1 - 1) Y[1]_n + e_n$$

3. **時間変化(historical change)**

$$Y[2]_n = \beta_0 + \beta_1 Y[1]_n + e[2]_n$$
$$Y[1]_n = \beta_0 + \beta_1 Y[0]_n + e[1]_n$$
$$(Y[2]_n - Y[1]_n) = \beta_1 \Delta Y[1\text{-}0]_n + \Delta e[2\text{-}1]_n$$

McArdle (2006) より作成

(b) 繰り返し測定されるデータに差分（difference-score）を用いるモデル
（Calculated "Difference-Score" Models for Repeated Measures）

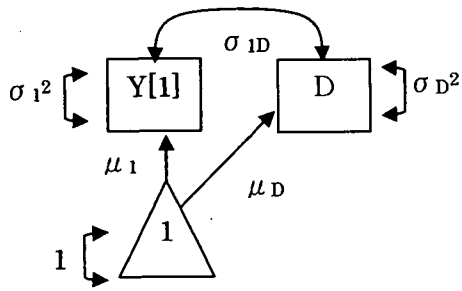
このモデルは，(a)のモデルと同様に測定されるイベントは繰り返し測定されるが，その推定値に差分（Y[2]-Y[1]）を加えて推定するモデルである。差分をD_nとすると以下のようになる。

$$Y[2]_n = Y[1]_n + D_n$$
$$Y[2]_n - Y[1]_n = D_n$$

差分の平均と分散，Y[1]との共分散は以下のようになる。

$$D_n = \mu_D + D_n^*$$
$$E\ \{D_n^{*2}\} / (N\text{-}1) = E\ \{D^* D^*\} = \sigma_D^2$$
$$E\ \{Y[1]^* D^*\} = \sigma_{1D}^2$$

これをパス図で表すと以下のようになる。



差分を加えることによって，2 時点における真の値（true score）がわかるため；2 時点における真の増加分がわかることになる。

(c) 繰り返し測定されるデータに潜在的な差分（"latent" difference score）を用いるモデル
("Latent" Difference-Score Models for Repeated Measures)

このモデルは，(a)のモデルと同様に測定されるイベントは繰り返し測定されるが，その推定値に潜在的な差分 Δy_D を加えて推定するモデルである。

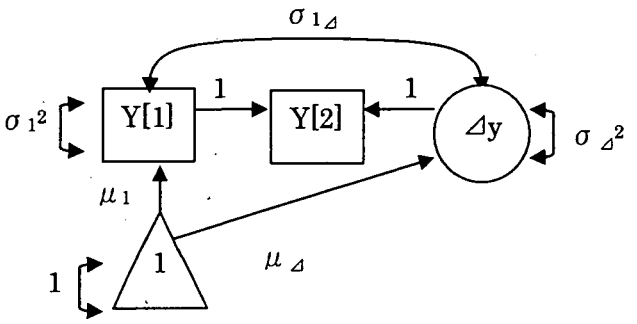
$$Y[2]_n = Y[1]_n + \Delta y_D$$
$$\Delta y_D = Y[2]_n - Y[1]_n$$

差分の平均と分散，Y[1]との共分散は以下のようになる。

$$\Delta y_D = \mu_d + \Delta y_D^*$$
$$E\ \{\Delta y_D^{*2}\} / (N\text{-}1) = E\ \{\Delta y^*\ \Delta y^*\} = \sigma_\Delta^2$$
$$E\ \{Y[1]^*\ \Delta y^*\} = \sigma_{1\Delta}^2$$

※ Δy は観察されないため，プロットができない。代わりに，観測値から統計的情報を用いて差分の情報を予測する。

これをパス図で表すと以下のようになる。⊿y は観測されないため，平均・分散・標準偏差は Y[2]=Y[1]+ ⊿y によって予測される。



潜在的な差分を用いるモデルにおいても，(a)や(b)と同様の情報を用いている。しかしこのモデルでは差分を直接算出しない分，系統的変化（systematic change）から得られる測定誤差を分離した値を得ることができるのである。

(d) 繰り返し測定されるデータを用いるモデルの要約
(Summary of Repeated Measures Models)

2 時点で繰り返し測定されるデータを用いて分析する場合，(a)~(c)のモデルをモデル適合度（goodness of fit tests）によって区別することは困難である。しかし差分を加えるなどモデルの係数変化の解釈によって漸く区別が可能である。測定回数が増えるにつれてそれぞれのモデルの差がみられるようになる。

4．所属集団の情報を加えた共分散構造分析
(Section D: Two-Occasion SEM with Group Information)

4－1．所属集団の情報を加えたマルチレベルモデル（Multi-Level models）

ここでは 2 時点で繰り返し測定されるモデルに所属集団の情報を加えたマルチレベルモデルについて説明する。マルチレベルモデルとは，ミクロ水準であるミクロデータ（個票データ）にマクロ水準である所属集団などの「階層的にネストされたデータ」（Kreft and Leeuw 1998[小野寺編訳 2006]）を分析するモデルである。「階層的に異なった水準（レベル）で測定された変数を含む解析モデル」（同上）ということでマルチレベルモデルと呼ば

れる。マルチレベルモデルは，「各々の文脈に対して別々（第1水準）の線形モデルを当てはめ，（中略）モデルは第2水準に組み込まれ，第1水準の回帰係数は第2水準の説明変数で回帰される」（同上）ものとして係数は解釈される。

所属集団の情報を加えた政経モデルは以下のようになる（n=1 to N）。

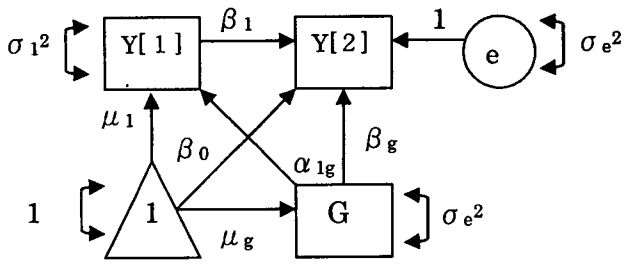
$$Y[2]_n = \beta_0 + \beta_1 Y[1]_n + \beta_g G_n + e_n$$

ここで，Gは2値変数である（「所属している」または「所属していない」）。
もし，Gがダミー変数（0,1）である場合，以下のようになる。

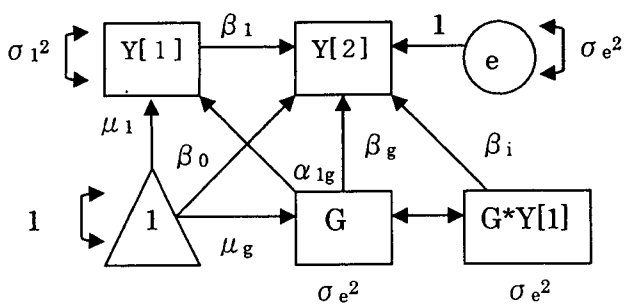
$$Y[2]_n\,[:G_n=0] = \beta_0 + \beta_1 Y[1]_n + \beta_g 0_n + e_n$$
$$Y[2]_n\,[:G_n=1] = \beta_0 + \beta_1 Y[1]_n + \beta_g 1_n + e_n$$

ここでG=0である場合，β_0は切片，β_1は傾きを表す。そしてG=1である場合，β_gは切片における変化を示す。

これをパス図で示すと以下のようになる。

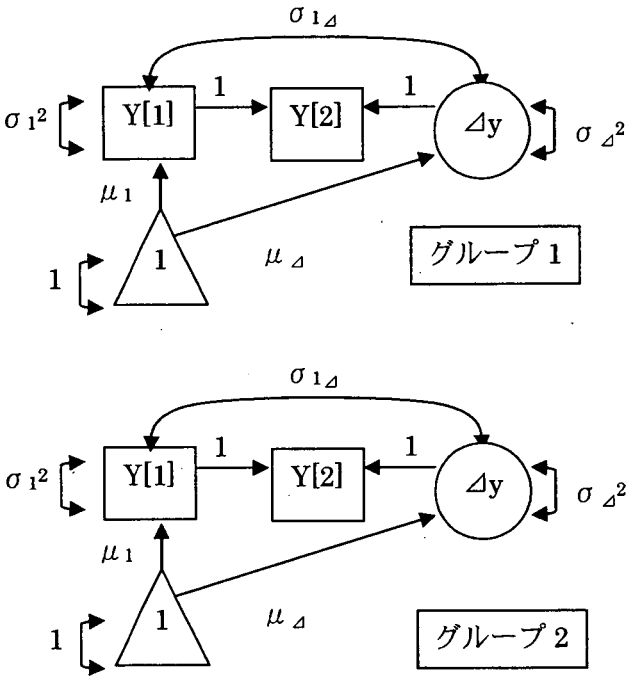


さらに，所属集団の情報とY[1]との交互作用項G*Y[1]を加えることで，所属集団の傾きの変化を示すことができる（β_i）。



4－2．複数の所属集団の情報を加えたモデル
(Adding Group Information in Multiple Groups MG-SEM)

これまでは，所属集団が 1 つで所属しているか否かのダミーのケースを考慮したが，実際は2つ以上の集団が存在するのが一般的である。このような場合のモデルを取り上げる。もし集団数が少ないとき（G＜10）は名義尺度（nominal categories）のカテゴリーとして扱う。その上で，集団間に差があるかどうかの検定を行い，差が存在する場合は共分散や平均を集団に属する個票に当てはめマルチレベルモデルを構築する方法や，集団ごとにモデルを分けて別々に推定する方法がある。



5．不完全なパネルデータを用いた場合の対策
(Section Q: Dealing with Incomplete Data in Longitudinal Studies)

不完全なパネルデータ（"missing" data）は単純なモデルであってもバイアスのかかった不正確な推定値を導きだしやすい。これはパネル分析に限らず多くの科学分野においてみられることである。このような不完全なデータを取り扱う際に，最も古典的な解決法としては完全なパネルデータ（balanced panel data）のみで分析を行う方法や欠損値部分を補完して行う方法がある。変数によっては欠損値を生みやすい特性を持つ場合があるため，単純に欠損値を除くだけでは，むしろ新たなバイアスを生み出す可能性も否定できないため，後者の補完の精度を上げる努力がより重要である。とはいえ，前者の完全なケースのみで行う分析を用いる場合が多い。具体的な処置を以下にまとめる。

- **削除（Deletion）**：ケースワイズ法（casewise，不完全なデータを全て削除する方法）やペアワイズ法（pairwise，分析に用いる変数群に不完全なデータが存在するときに対象サンプルを削除する方法）によって完全なデータを用いて適用する方法である。処理は単純（simplicity）で明確（clarity）で広範囲（wide-spread）な方法であることが期待され，サンプルのロス，標準誤差の増大，欠損値がランダムでない場合にバイアスが生じるといった注意が必要である。
- **重み付け（Weighting）**：バイアスを修正するような重み付けを施してデータに適用する方法である。
- **修正，補完（Imputation）**：データの情報をもとに欠損値（"missing" data）や平均値を修正，補完する方法である。完全なデータを用いて算出した平均を代入（"mean substitution"）し，再推定を行う方法である。これによって真の平均は失われることや不正確な標準誤差や自由度，バイアスがかかることに注意する必要が生じる。他には回帰モデルによって予備推定を行い誤差分散の推定値を用いてランダムに不完全なデータを代入する方法もある。

McArdle (2006) より作成

データが不完全（"incompleteness"）であることは，欠損値のパターンがどのような特性を持つかを十分に明らかにした上で，対策を決める必要がある（完全なデータのみを用いるのか，重み付けや補完を行うのか）。単純な方法としては，欠損値が存在するデータとそうでない完全なデータで欠損値の存在しない他の変数の記述統計を算出して比較する方法がある。この方法の目的は，完全なデータと不完全なデータの差があるかどうかという点よりもどのぐらいの差が生じているかを詳細に記述，観察することである。その上で，サンプリングによって得られた期待値よりも利用できるデータはどの程度異なるのかをみるのである。

6．カテゴリカル・データを用いた共分散構造分析 (Section S: Longitudinal Analysis with Categorical Outcomes)

社会科学における調査データにおいて，カテゴリカル・データは最も一般的な変数である。カテゴリカル・データは一般的に正規分布を仮定するモデルが多い中で，情報が少ない。統計的な問題点は 2 値変数（binary measures）のような制限された情報をどのように扱うかによるものである。

カテゴリカル・データにおける項目は response propensity（response strength）と呼ば

れる潜在変数によって規定し，正規分布に従うと仮定する。その上で，カテゴリーを区分する境界点（threshold）を仮定する。もし response propensity が境界点よりも大きければ（小さければ），各サンプルの回答が正の値（負の値）となる。境界点は多くの物理現象において使用される一般的なモデルである（図 5）。

変数のカテゴリーの数は多ければ多いほど連続変数に近くなるという点でバイアスは小さくなる。その点で 2 値変数は大きなバイアスをもった変数であるということがいえる。このような欠点を補うために，擬似的に連続変数化させる方法がある。それが，テトラコリック相関係数（Tetrachoric correlation coefficient）である。テトラコリック相関係数は以下のように定義される。

$$r \fallingdotseq \cos\left[\frac{\pi\sqrt{bc}}{\sqrt{ad}+\sqrt{bc}}\right]$$

	0	1	
0	a	b	
1	c	d	

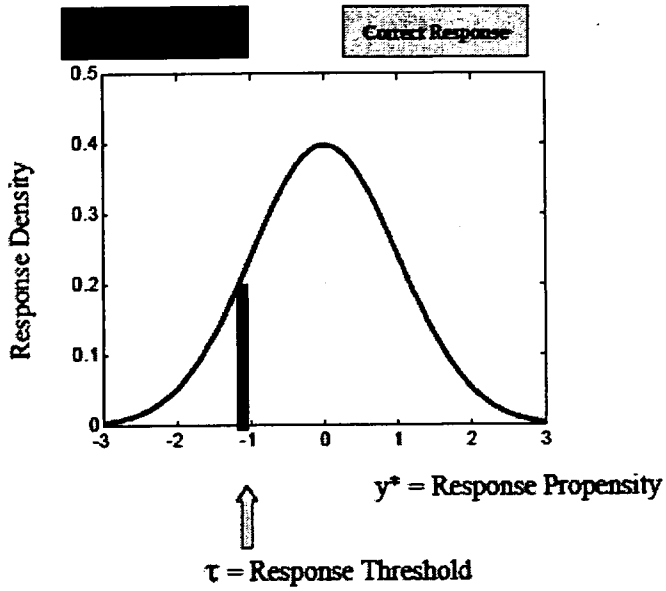
ここでは 2 値変数を用いた場合についてレビューする。測定される 2 値変数が繰り返し測定される場合を想定し，ロジスティック回帰モデルを適用する場合を取り上げる。ロジスティック回帰モデルは以下のようになる。P(g) / (1-P(g))はオッズを示す。

$$\ln\{P(g)/(1-P(g))\} = B_0 + B_1{*}X(g)$$

このモデルをオッズとイベントの発生確率について表すと，

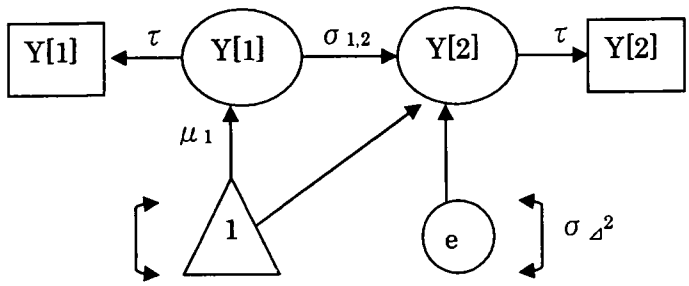
$$\{P(g)/(1-P(g))\} = \exp\{B_0 + B_1{*}X(g)\}$$
$$P(g) = \exp\{B_0 + B_1{*}X(g)\} / (1+\exp\{B_0 + B_1{*}X(g)\})$$

図 5　Response propensity の仮定と境界点の関係



Hamagami & McArdle (2005)より転載

このモデルは，最尤推定法を用いて推定される。パス図で示すと以下のようになる。τ は観測された 2 値変数（四角で囲まれた Y[1]と Y[2]）を擬似的に連続変数に変換された境界点を示している。



カテゴリカル・データを用いた場合の構造方程式モデルは擬似的に連続変数に変換すること以外の構造は他のモデルと同様の解釈が可能である。

参考文献

McArdle J. John and Fumiaki Hamagami, 2006. "Structural Equation Modeling in Longitudinal Research", Longitudinal Research Institute Workshop.

Section B, Section Q, Section S.

Hamagami Fumiaki, 2005. "Multiple Time Categorical Factor Models", APA-LRI Workshop.

小野寺　孝義(編訳)，岩田　昇(訳)，菱村　豊(訳)，長谷川　孝治(訳)，村山　航(訳)，2006.『基礎から学ぶマルチレベルモデル』，ナカニシヤ出版．Kreft Ita and Jan de Leeuw, 1998 "Introducing Multilevel Modeling", Sage Publications Ltd.

竹内　啓（監修），豊田　秀樹（著者），1992.『ＳＡＳによる共分散構造分析　ＳＡＳで学ぶ統計的データ解析③』，東京大学出版会.

（2） パネル調査の統計分析モデル：マルチレベルモデルと適用例

鎌田 健司

1．はじめに

本稿では，パネル調査の統計分析手法としてマルチレベルモデル（Multilevel Model）を取り上げる。マルチレベルモデルは，それ自体が固有の分析手法として成り立っているのではなく，回帰分析などの分析に新たな視点を与えるモデルの 1 つである。そのため，イベントヒストリー分析や共分散構造分析において用いることが可能である。一分析モデルなので，パネル分析に特化した分析手法というわけではないが，適用できる分析手法が多いことに特長がある。

マルチレベルモデルは，ミクロ水準であるミクロデータ（個票データ）にマクロ水準である所属集団などの「階層的にネストされたデータ」（Kreft and Leeuw 1998，小野寺編訳 2006）を組み込んで分析するモデルである。「階層的に異なった水準（レベル）で測定された変数を含む解析モデル」（同上）ということでマルチレベルモデルと呼ばれる。ただし，想定される変数間の連関や仮説設定，データ構造によって，成長曲線モデル（Growth Curve Model），階層線形モデル（Hierarchical Linear Model），一般化線形混合モデル（Generalized linear mixed model）などとも呼ばれる。

マルチレベルモデルは，「各々の文脈に対して別々（第 1 水準）の線形モデルを当てはめ，（中略）モデルは第 2 水準に組み込まれ，第 1 水準の回帰係数は第 2 水準の説明変数で回帰される」（同上）ものとして係数は解釈されるモデルであるということができる。例えば，地域要因（マクロ水準）を個人要因（ミクロ水準）と同時に推定する場合，地域要因の影響はその地域に所属するすべての個人に共通する性質として地域要因と個人要因は相関をもつため，最小二乗法（OLS）の仮定である変動項の独立を満たすことができなくなる。これにより，推定値の標準誤差は実際よりも小さく推定されるため，独立変数の係数は統計的に有意になりやすくなるという誤差が生じる。マルチレベルモデルによって推定することによって，地域要因によって個人要因の従属変数への影響が異なる場合に，前述の誤差の問題点を解決することができる。

具体的には，最も単純なマルチレベル分析においては，従属変数の地域差を加味した分析を行う場合，マクロレベルの数値をミクロレベルの所属指標の平均値で代替することによってマクロ要因とミクロ要因の推定値を分離して推定することができる。

マルチレベルモデルの適用上の制約については，マクロ水準のカテゴリー数が少なく，カテゴリーに含まれるミクロ水準のサンプルが極端に多い場合はマクロ要因とミクロ要因の推定値の分離が困難になるという点である（Goldstein 1999）。

2．データ構造

マルチレベルモデルが適用可能なデータ構造は以下の2つがある（Painter 2006）。1つは対象サンプルを 1 レコードとし，ミクロ要因とマクロ要因に関する変数を列挙する方法で図となっている（Multiple-Variable Structure：MV 構造）。もう1つは，繰り返し従属変数が集計されるデータにおいて，集計ごとにレコードが追加されるデータ構造で図となっている。対象サンプルの ID 変数を用いて管理する方法である（Multiple-Recode Structure：MR 構造）。基本的なマルチレベルモデルにおいては MV 構造のデータを用いて推定することができ，MR 構造のデータはイベントヒストリー分析を行う場合など時間経過やタイミングを考慮する場合に適用することができる。

図1　データ構造：(a) MV 構造，(b) MR 構造（Painter 2006，pp. 2-3）

(a)　Mutiple-Variable 構造

ID	Var1	Var2	Var3	Group
1	12	45	34	A
2	23	43	34	B
3	31	54	45	C
4	13	42	31	A
5	26	40	38	B
6	27	49	44	C

(b)　Mutiple-Recode 構造

ID	Var X	Order
1	12	1
1	45	2
1	34	3
2	23	1
2	43	2
2	34	3
3	31	1
3	54	2
3	45	3

3．マルチレベルモデルの基本概念

ここでは，マルチレベルモデルに用いる基本概念を簡単にまとめる。

3－1．文脈モデル contextual model（Kreft and Leeuw 1998，小野寺編訳 2006, p.7, 19-30）

文脈モデルはマルチレベルモデルの最単純モデルの1つである，2つのタイプの変数を含む回帰モデルである。2つのタイプとは，ミクロ水準の変数とマクロ水準を平均や中央値のように集計した文脈変数を示す。集計された文脈変数を含む回帰分析を文脈モデルという。低水準の変数と，集計されたり，全体として測定されたりした高水準の特徴を含むすべての線形回帰モデルである。

文脈モデルの線形モデルは以下の通りである。

$$y_{ij}=a+bx_{ij}+cz_j+e_{ij} \qquad (1.1)$$