2　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. "Structural Equation Modeling in Longitudinal Research", Longitudinal Research Institute Workshop 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。

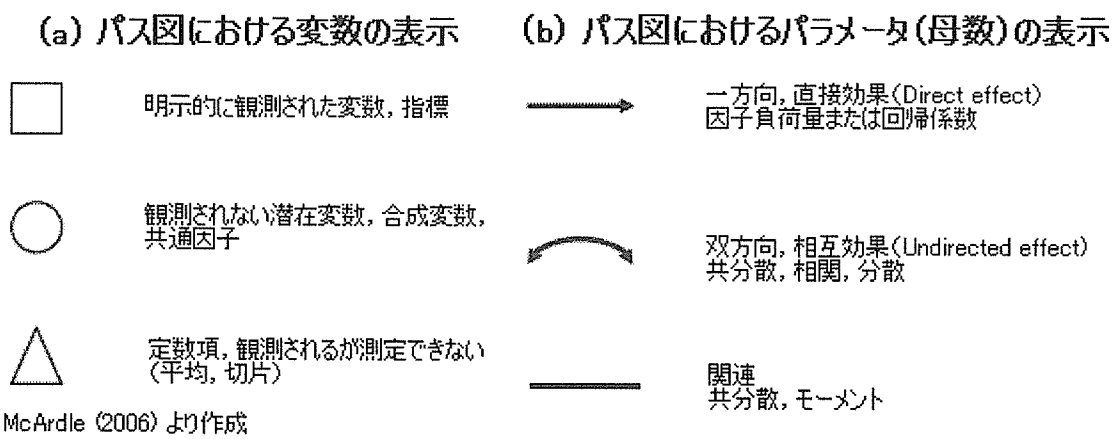
2．共分散構造分析の基礎概念

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。

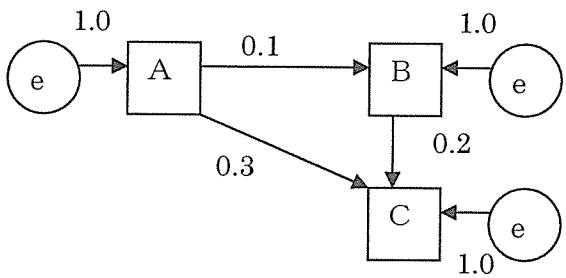
図 1 は構造方程式をグラフ化（パス図）した場合の表示形式を示したものである。四角は明示的に観測された変数を示し，丸は観測されない潜在変数，合成変数，共通因子を示し，三角は観測されるが測定できない変数として定数項（変数，切片）を示す。パラメータの表示については，一方向の矢印は直接効果（direct effect）を示し，推定方法によって回帰係数や潜在変数を想定した場合の因子負荷量を表す。双方向の矢印は相互効果（Undirected effect）を示し，推定方法によって共分散，相関，分散を表す。矢印無しの棒線は変数間の関連を示し，推定方法によって共分散，モーメントを表す。

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

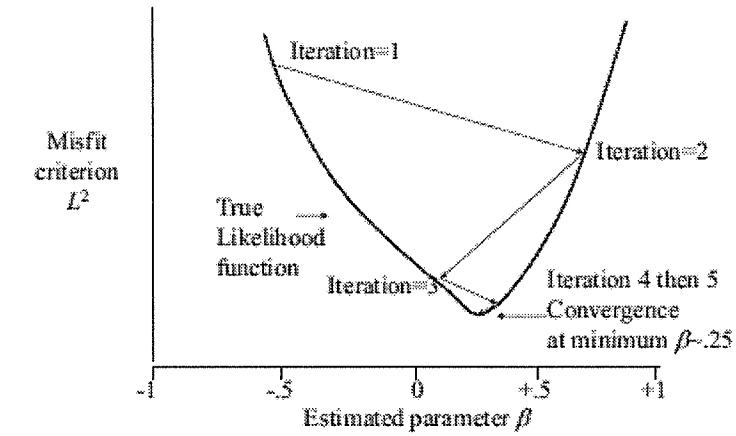
3－1．共分散構造モデルの分析過程

（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）。

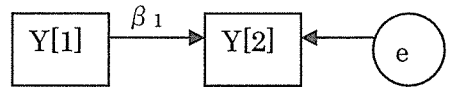
３－２．２時点における構造方程式モデル（two-occasion longitudinal data with SEM）

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。

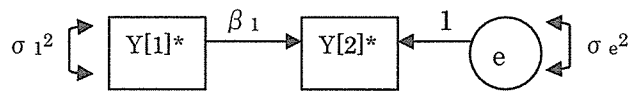
(a) 繰り返し測定するデータを用いた自己回帰モデル
（A typical auto-regression path model for two repeated measures）

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。

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



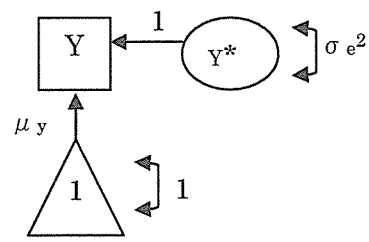
次に共分散を用いた自己回帰モデルを想定すると以下のようになる。アスタリスク（*）は平均周辺の偏差を示す。



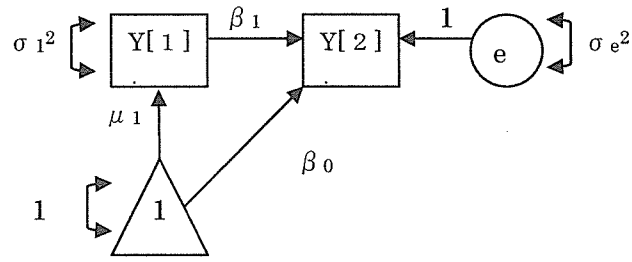
ここで，Y に対する平均と分散は以下のように定義される。μ_0はグループ内の定数項を示し，Y_n*は各サンプルの平均周辺の偏差（$Y_n - \mu$）を示す。

$$Y_n = \mu_y + Y_n^*$$

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



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）

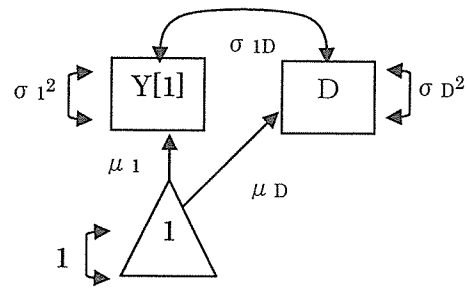
このモデルは，(a)のモデルと同様に測定されるイベントは繰り返し測定されるが，その推定値に差分（Y[2]-Y[1]）を加えて推定するモデルである。差分をD_nとすると以下のようになる。

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

差分の平均と分散，Y[1]との共分散は以下のようになる。

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

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。

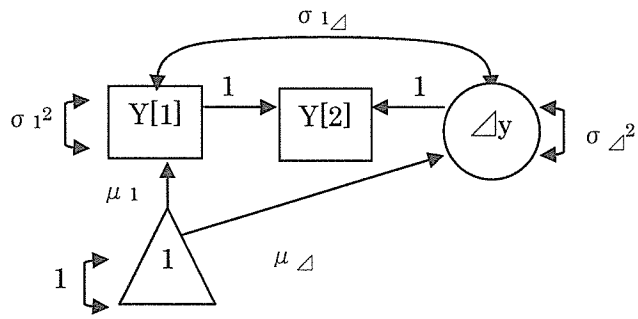
$$Y[2]_n = Y[1]_n + \Delta y_D$$
$$\Delta y_D = Y[2]_n - Y[1]_n$$

差分の平均と分散，Y[1]との共分散は以下のようになる。

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

※ 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)。

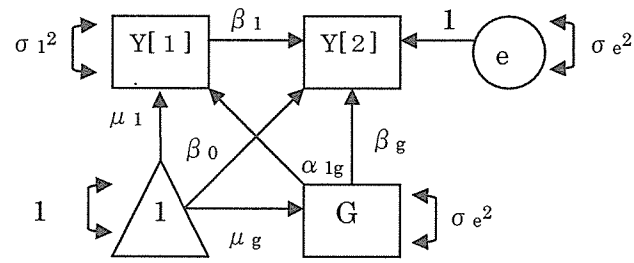
$$Y[2]_n = \beta_0 + \beta_1 Y[1]_n + \beta_g G_n + e_n$$

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。

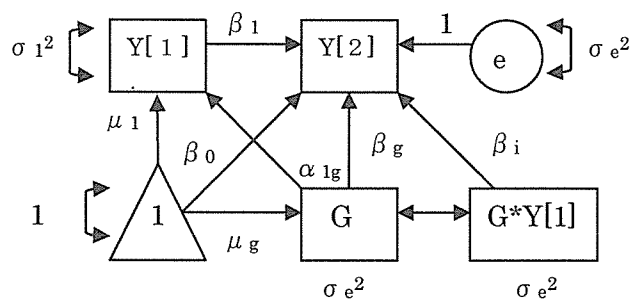
$$Y[2]_n\ [:G_n=0] = \beta_0 + \beta_1 Y[1]_n + \beta_g 0_n + e_n$$
$$Y[2]_n\ [:G_n=1] = \beta_0 + \beta_1 Y[1]_n + \beta_g 1_n + e_n$$

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。

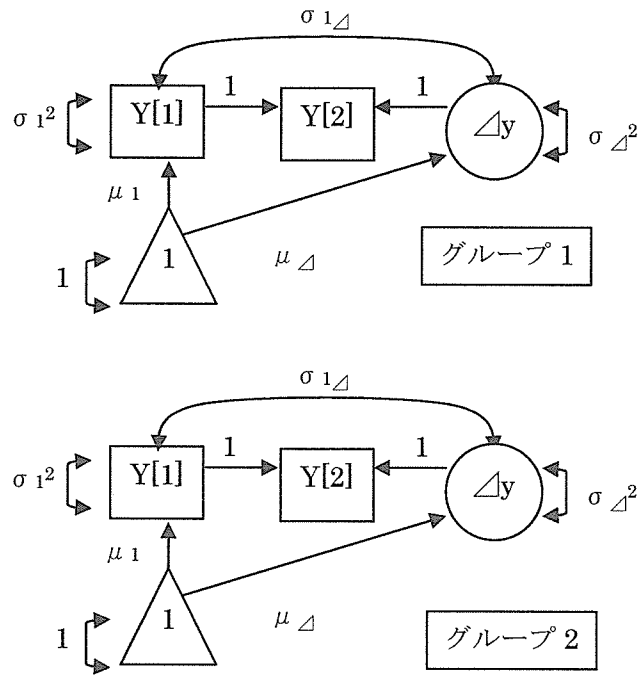


さらに，所属集団の情報と Y[1]との交互作用項 G*Y[1]を加えることで，所属集団の傾きの変化を示すことができる（β_i)。



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	0
1	c	d	1

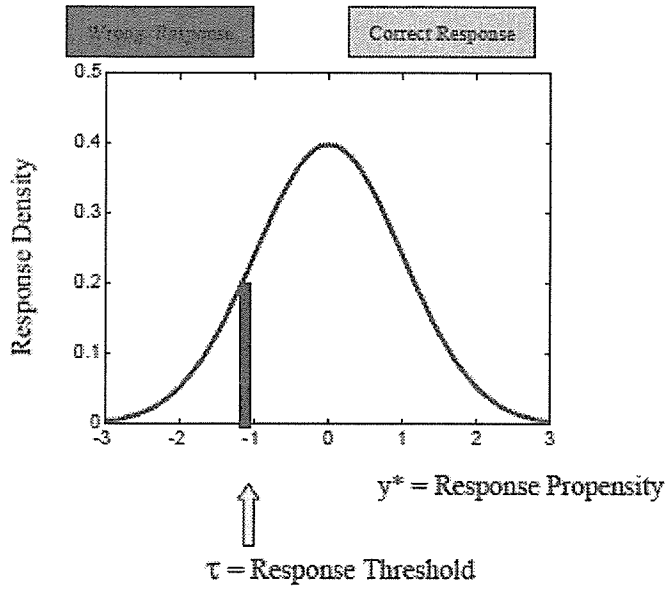
ここでは 2 値変数を用いた場合についてレビューする。測定される 2 値変数が繰り返し測定される場合を想定し，ロジスティック回帰モデルを適用する場合を取り上げる。ロジスティック回帰モデルは以下のようになる。P(g) / (1-P(g))はオッズを示す。

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

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，

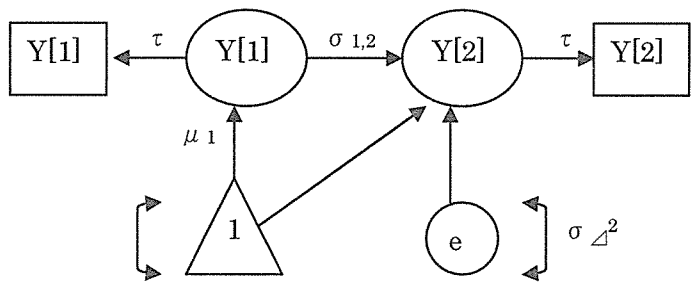
$$\{P(g)/(1-P(g))\} = \exp\{B_0 + B_1{*}X(g)\}$$
$$P(g) = \exp\{B_0 + B_1{*}X(g)\}/(1+\exp\{B_0 + B_1{*}X(g)\})$$

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．

3　第1回～第4回21世紀出生児縦断調査の脱落・移動の動向

西野 淑美

21世紀出生児縦断調査の第3回までの脱落・移動の動向について、筆者は昨年、『パネル調査(縦断調査)のデータマネジメント方策及び分析に関する総合的システムの開発研究』(厚生労働科学研究費補助金)平成17年度報告書所収の論文（西野、2006）でまとめた。本稿では、第4回までの動向を簡単に確認したい。

1．脱落・移動の全体動向

第1回調査の回答者のうち、第4回調査に回答した人は88.4%である。図表1に示したように、全体の84.7%は第1-4回全てに回答しており、途中に抜けた回があるが第4回は回答している人が全体の3.7%である。

図表1　第1回～第4回の脱落状況のまとめ

		度数	ﾊﾟｰｾﾝﾄ	有効ﾊﾟｰｾﾝﾄ	累積ﾊﾟｰｾﾝﾄ
有効	第1-4回全て回答	39839	84.7	84.7	84.7
	第2回以降全て脱落	1943	4.1	4.1	88.9
	第3回以降全て脱落	1318	2.8	2.8	91.7
	第4回のみ脱落	1826	3.9	3.9	95.6
	第1回回答、第2回脱落、第3･4回復活回答	778	1.7	1.7	97.2
	第1･2回回答、第3回脱落、第4回復活回答	942	2.0	2.0	99.2
	第1回回答、第2回脱落、第3回復活回答、第4回脱落	369	.8	.8	100.0
	合計	47015	100.0	100.0	

この中で、復活回答に注目したい。第2回調査に回答していなかった（脱落していた）人のうち、37.1%の人は、第3回で再び回答（復活回答）を寄せている。第3回調査に回答していなかった人の22.4%は、第4回調査には回答している。転居先不明で郵便が届かない場合でない限りは、一度脱落した人にも調査票を届ける努力を続けることは、決して無駄ではない。

移動（居住自治体の変化）については、第1回→第2回が8.2%（第2回の脱落者を除いて算出、以下同）、第2回→第3回が7.3%、第3回→第4回は8.9%の回答者に生じていた。第1回から第4回の間で1度でも移動をしたケースは、全体の22.8%にのぼる（脱落で判別できないケースを除く）。ただし、移動はあくまで、調査対象者が住所変更を調査事務局に届けた場合に判明していることに留意する必要がある。また、昨今の自治体合併に伴い、自治体名の変更なのか自治体移動なのか判別できないケースが、各回とも0.5-1.6%いることも報告する。

2．第4回回答サンプルと理想のサンプルとの比較—脱落による歪みの検証—

第 1 回調査に回答した全サンプル、すなわち脱落なしで全員が継続回答した場合という「理想」のサンプルと、脱落せずに実際に第 4 回調査まで継続回答したサンプルとで、第 1 回調査の諸変数の値を比較した。また、同じく、第 1 回調査全サンプルと、途中脱落して復活した人も含めて第 4 回に回答した全サンプルとを比較した。いずれもとりあげたのは、昨年度の分析（西野、2006:195-196）で何らかの有意差があった項目である。

具体的には、第 1 回調査の全サンプルを母集団として想定し、離散変量は χ^2 検定で、連続変量は t 検定で、1 サンプルによる検定を行うことで、脱落によるサンプルの歪みが生じているかを確認した。有意差が生じているとしたら、その分実際の回答者のサンプルは、全員が継続回答した「理想の」サンプルと比べて、歪みが生じていることになる。この方法は、12 年間の高齢者縦断調査の回答者と脱落者の特性比較および、脱落の無い理想のサンプルと現実のサンプルについて初回調査での各変数の値の比較を行った（杉澤他、2000）を参考にしている。いずれかの年度に一つでも有意差があった変数について、図表 2 にまとめた。なお、2 回目・3 回目の値は、昨年の筆者の分析と同じものである（ただしミスが見つかった部分は修正した）。

分析の結果からは、第 4 回の実際のサンプルは、理想サンプルと比べて、母親・父親の年齢が若いケース、収入が低いケース、父母のどちらかが外国人であるケース、6 ヶ月の時点でひとり親のケースや父・母がふだんの保育に関わっていないケース、職・収入・育児・家事・相談相手などで父親のプレゼンスが低いケース、6 ヶ月時に保育士や保育ママ等を利用していたケース、悩みを相談する相手がいない人、配偶者・両親・友人知人・保健師が相談相手になっていないケース、6 ヶ月までに今回の妊娠出産に伴う引越（増築）があったケース、喫煙本数が多いケースが、抜け落ちる方向で歪んでいると解釈できる。また、ふだんの保育者の組み合わせが「親と祖父母」のケース、人工乳を使わなかったり母乳の授乳期間が長かったケースが多くなる方向で歪んでいると考えられる。これらの歪みは、ほとんどの項目で前回より拡大している。また、これらの傾向は、復活者を含んだ場合も含まない場合もほぼ同じである。

ひとり親、若い、外国籍、父親のプレゼンスが低い、相談相手がいないなど、いずれも支援を必要とする可能性が高いサンプルが抜け落ちる傾向にあることは、21 世紀出生児縦断調査の結果を政策に応用していく際には、留意する必要がある。また、引越ケースの脱落についても有意であることは、移動者の追跡が必要であることを物語る。今後も、脱落傾向の観察を続け、得られた結果に応じて類似調査の企画の際には何らかの対策を練っていくことが重要であろう。

【引用文献】

西野淑美 (2006)「21 世紀出生児縦断調査における脱落・居住地移動・復活サンプルの分析」『パネル調査(縦断調査)のデータマネジメント方策及び分析に関する総合的システムの開発研究』(厚生労働科学研究費補助金)平成 17 年度報告書、p181-207.

杉澤秀博他 (2000)「全国高齢者に対する 12 年間の縦断調査の脱落者・継続回答者の特性」『日本公衆衛生雑誌』47(4):337-349。

図表2　脱落の有無により変数がとる値及び第1回調査との有意差の有無

		第1回	第2回	第3回		第4回	
集計対象		回答者	回答者	回答者	回答者(復活除く)	回答者	回答者(復活除く)
除いた対象		なし	脱落者	脱落者	脱落者・復活者	脱落者	脱落者・復活者
2000年12月31日時点での父親年齢	平均値	31.26	31.37 ***	31.42 ***	31.45 ***	31.45 ***	31.51 ***
2000年12月31日時点での母親年齢	平均値	29.08	29.21 ***	29.25 ***	29.30 ***	29.31 ***	29.37 ***
父母とも日本人	(該当;%)	96.6	97.0 ***	97.1 ***	97.2 ***	97.2 ***	97.3 ***
同居の状況(母)	(同居;%)	99.9	99.9	99.9 *	99.9 *	99.9 *	99.9 **
同居の状況(父)	(同居;%)	97.7	97.9 ***	98.0 ***	98.1 ***	98.1 ***	98.2 ***
同居の状況(母の母親)	(同居;%)	6.4	6.3	6.2 *	6.1 *	6.1 *	6.0 **
兄姉弟妹の人数(双子込み)	平均値	0.69	0.68	0.68	0.68 *	0.68	0.68 *
第1回　核家族世帯	(該当;%)	76.7	76.9	77.0	77.1 *	77.0	77.2 *
第1回　三世代世帯	(該当;%)	20.5	20.6	20.6	20.5	20.6	20.6
第1回　ひとり親世帯(祖父母同居含む)	(該当;%)	2.3	2.0 ***	2.0 ***	1.9 ***	1.9 ***	1.8 ***
妊娠出産に伴う引越・増築の有無	(あり;%)	11.7	11.3 **	11.2 ***	11.1 ***	11.1 ***	11.0 ***
ふだんの保育者(母)	(該当;%)	97.1	97.3 *	97.3	97.3 **	97.3 *	97.4 **
ふだんの保育者(父)	(該当;%)	46.4	47.1 **	47.2 ***	47.5 ***	47.5 ***	47.9 ***
ふだんの保育者(祖母)	(該当;%)	20.9	21.0	21.0	21.0	21.1	21.1
ふだんの保育者(保育所の保育士)	(該当;%)	3.9	3.7	3.7 *	3.6 *	3.6 *	3.6 **
第1回　ふだんの保育者　親と祖父母	(該当;%)	19.9	20.0	20.1	20.1	20.9 ***	20.2 ***
第1回　ふだんの保育者　親と保育士等	(該当;%)	2.0	2.0	1.9	1.9 *	2.0	1.9
保育士や保育ママやベビーシッターの利用	(あり;%)	4.2	4.1	4.0 *	4.0 *	4.0 *	3.9 **
授乳は母乳のみ	(該当;%)	21.0	21.5 *	21.6 **	21.8 ***	21.7 ***	21.8 ***
母乳を与えた期間(月)	平均値　与えていない=0ヶ月	4.48	4.5 ***	4.6 ***	4.6 ***	4.6 ***	4.6 ***
父の育児休業取得期間(月)	平均値　有職だが利用無=0ヶ月	0.012	0.012 ***	0.011 ***	0.011 ***	0.012 ***	0.011 ***
母の家事(食事をつくる)　4段階(1～4)	平均値　1に近いほど頻度が高い	1.09	1.08 *	1.08	1.08	1.08 *	1.08 *
父の育児(入浴させる)　4段階(1～4)	平均値　1に近いほど頻度が高い	1.81	1.80	1.80	1.80 *	1.80	1.80 *
子を持ってよかったこと(身近な人が喜んでくれた)	(該当;%)	78.1	78.4	78.5	78.6 *	78.5	78.7 **
子を持って負担に思うこと(子育てによる身体の疲れが大きい)	(該当;%)	39.5	39.8	39.8	40.0 *	39.9	40.0 *
子を持って負担に思うこと(子育てで出費がかさむ)	(該当;%)	34.7	34.4	34.2 *	34.2 *	34.1 **	34.0 **
子を持って負担に思うこと(自分の自由な時間が持てない)	(該当;%)	55.2	55.6	55.7 *	55.8 **	55.8 **	56.0 **

図表2(つづき) 脱落の有無により変数がとる値及び第1回調査との有意差の有無

		第1回	第2回	第3回		第4回	
集計対象		回答者	回答者	回答者	回答者(復活除く)	回答者	回答者(復活除く)
除いた対象		なし	脱落者	脱落者	脱落者・復活者	脱落者	脱落者・復活者
子育ての不安や悩みを相談する人	(あり;%)	99.0	99.1	99.1 *	99.2 *	99.2 *	99.2 **
子育ての相談相手(配偶者)	(該当;%)	81.5	82.4 ***	82.5 ***	82.8 ***	82.9 ***	83.2 ***
子育ての相談相手(自分の両親)	(該当;%)	72.3	72.6	72.8 *	72.9 **	73.0 **	73.2 ***
子育ての相談相手(配偶者の両親)	(該当;%)	30.3	30.5	30.7 *	30.8 *	30.9 *	31.0 **
子育ての相談相手(友人・知人)	(該当;%)	70.5	70.9	71.0 *	71.1 **	71.1 **	71.3 ***
子育ての相談相手(保健師)	(該当;%)	14.2	14.4	14.6 *	14.6 **	14.6 *	14.8 **
出産1年前の父の職の有無	(あり;%)	98.3	98.5 *	98.5 **	98.6 ***	98.6 **	98.6 ***
月齢6ヶ月時の父の職の有無	(あり;%)	98.3	98.4	98.5 *	98.5 **	98.5 **	98.5 **
母の労働時間 5段階(1～5)	平均値 1に近いほど短い	1.27	1.26	1.27	1.26 *	1.27	1.26 *
母の通勤時間 6段階(0～5)	平均値 0に近いほど短い	0.25	0.25	0.25	0.25	0.25	0.25 *
父の労働時間 5段階(1～5)	平均値 1に近いほど短い	4.02	4.03 *	4.03 **	4.04 ***	4.04 ***	4.04 ***
父の通勤時間 6段階(0～5)	平均値 0に近いほど短い	2.45	2.46 **	2.46 **	2.47 ***	2.47 **	2.47 ***
母の就労収入(有無)	(あり;%)	50.7	50.9	51.0	51.0	48.8	48.8 *
母の就労収入(金額:万円)	平均値	99.4	101.1 *	101.6 **	102.1 ***	102.5 ***	103.1 ***
父の就労収入(有無)	(あり;%)	98.8	99.0 **	99.0 **	99.0 ***	99.0 ***	99.1 ***
父の就労収入(金額:万円)	平均値	445.5	451.1 ***	453.0 ***	455.0 ***	455.1 ***	457.7 ***
父母の就労収入(有無)	(あり;%)	99.4	99.5 *	99.5 *	99.5 **	99.5 **	99.5 ***
父母の就労収入(金額:万円)	平均値	546.1	553.3 ***	555.7 ***	558.1 ***	558.7 ***	561.7 ***
父母の就労収入＋その他の収入(有無)	(あり;%)	99.6	99.7	99.7	99.7 *	99.7 *	99.7 **
父母の就労収入＋その他の収入(金額:万円)	平均値	557.3	564.6 ***	567.3 ***	569.5 ***	570.3 ***	573.3 ***
父母の就労収入＋その他の収入(子育て費用の割合)	‰	80.5	78.3 ***	77.4 ***	76.7 ***	76.8 ***	75.6 ***
1ヵ月の子育て費用(子育て費用:万円)	平均値	4.1	4.0	4.0 *	3.9 **	3.9 **	3.9 ***
保育料(有無)	(あり;%)	5.8	5.6	5.6	5.5 *	5.6	5.5 *
母の1日の喫煙本数	平均値 吸わない=0本	2.01	1.8 ***	1.8 ***	1.7 ***	1.7 ***	1.6 ***
父の1日の喫煙本数	平均値 吸わない=0本	11.81	11.6 ***	11.5 ***	11.5 ***	11.5 ***	11.4 ***

4　縦断調査マイクロシミュレーション分析システムの設計・開発

金子　隆一
三田　房美

1．はじめに

パネル調査(縦断調査)の分析法の一つとして、マイクロシミュレーション分析がある。マイクロシミュレーションとは、各種属性を持った個人の集団をコンピュータ上に構成して、おのおのの行動や状態変化を発生させることにより、集団の変化を再現するシミュレーション技法である。対象集団の将来予測、行政制度・施策の効果の予見をはじめ、行動メカニズムの解明や統計手法の精度評価など、幅広く応用される。一方、パネル調査は、抽出された標本内の同一対象(個人、世帯)を追跡しながら継続的に調査し、対象者の変化とその要因を記録して行くものであり、その枠組みやデータ構造はマイクロシミュレーションにきわめて近く、両者はきわめて親和性が高いといえよう。実際、諸外国においては、社会政策、税制等の制度・施策の評価や検討のためにパネル調査に基づいたマイクロシミュレーション分析が行われている。

本事業では、21 世紀縦断調査の結果を元に、その分析対象となる結婚、出生、就業などの事象の発生メカニズム、決定要因の解明や、介入（たとえば制度・施策の実施）の効果の評価・予測を行う際に有力な分析手段となるマイクロシミュレーション分析を行うこととしている。また、パネル調査の統計分析上の弱点ともいえる標本脱落や回答不詳・不整合の影響を評価する方法ともなりうるので、既存の統計モデルの検証に用いることで、より信頼性の高い分析結果を提供できると考えられる。したがって、それら既存の統計モデルによる分析と合わせてマイクロシミュレーション分析を行うことによって、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。

本事業の先行事業に当たる研究では、諸外国のマイクロシミュレーション分析の事例について検討を行い、21 世紀縦断調査に特化したマイクロシミュレーション分析を行うための支援システムの設計を行い、開発に着手した。すなわち、パネル調査データの管理情報を基に、シミュレーション分析に必要となる標本モデルをシミュレーション言語（現行では C++）と連携しながら生成するシステムを作成した（システムは、本事業で構築を行ったデータマネジメントシステムの一環として開発されており、統合的に扱うことができるものである)。

本年度の事業では、これらを元にしてシステム開発の事業者とともに、実際のライフコースモデルを組み込んだマイクロシミュレーションモデルを開発、実装する予定であったが、予算的制約等によりこれを見送らざるを得なかった。これらは次年度以降に行われる予定である。したがって、本年度はその実現に向けて、設計を進め、またシステムの基礎的部分の強化に努めた。

2．縦断調査用マイクロシミュレーション分析システムの開発

マイクロシミュレーションとは、各種属性を持った個人の集団をコンピュータ上に構成して、おのおのの行動や状態変化を発生させることにより、集団の変化を再現するシミュレーション手法である。とくに縦断型マイクロシミュレーション longitudinal micro-simulation と呼ばれるものは、個人の経時的変化を模擬するもので、パネル調査データとの親和性が高く、対象集団の変化の将来予測、行政制度・施策の効果の予見をはじめ、行動メカニズムの解明や統計手法の精度評価など、既存の統計分析に止まらない多くの応用と可能性を持っている。パネル調査で捉えられた標本をシミュレーションモデルとして再現すれば、さまざまな仮想的条件や仮定の下での標本の変化を観察することが可能であり、それらを実際の変化と比較すれば、仮定やモデルの妥当性を評価することができる。

縦断型マイクロシミュレーションは、21 世紀縦断調査についても、その主要なテーマである結婚・出生・子育てなどの発生メカニズムと決定要因の解明や、制度・施策効果の評価を行う有力な手法となるほか、脱落をはじめとするパネル調査特有の統計分析上の困難に対して、さまざまな条件下におけるそれら統計手法の妥当性や精度を検証する有効な手段を与えると考えられる。

本研究では、21 世紀縦断調査データを活用して今後継続的なマイクロシミュレーション分析が行えるよう、その基礎としてエージェント型(agent-base)のマイクロシミュレーションモデルに必要な標本を生成するシステムを開発した。これはパネル調査データの管理情報を活用して、シミュレーション分析に必要となる標本モデルを半自動的に生成するシステムであり、現行では C++によるシミュレーションモデルを作成することができる。システムは、本事業で構築を行ったデータマネジメントシステムの一環として開発されており、統合的に扱うことができるものである。

(1) マイクロシミュレーションにおける標本モデル

ここで想定する縦断型マイクロシミュレーションは、エージェント型(agent-base)のマイクロシミュレーションモデルを基礎とするものである。そこでは個人のモデルは、自律性を備えたオブジェクト、すなわちエージェントして実装される。図１には、本シミュレーションのベースモデルとなるプロトタイプモデルのクラス図を示した。これは観察単位（エージェント）の時間的変化・行動を継続的に発生するタイプのクラスの定義である（クラスとは、エージェントまたはオブジェクトのシミュレーション言語上の定義のことである)。21 縦断調査の対象者に対応するエージェント・クラスを中心として、その属性や家族などの関係者、さらには標本集団とその統計的特性を集合的に計測、記録、出力する統計制度のエージェント・クラスを配置している。これらを基本とし、出生児調査、成年者調査など各調査ごとに、また分析テーマごとに、必要なエージェント・クラスを追加して分析モデルを構築することとなる。