第 7 章

共分散構造分析

本章では，縦断調査を用いた場合の共分散構造分析の適用について、いくつかのモデルを参照し、社会科学データで用いられることの多いカテゴリ変数を用いる場合の適用例について紹介する。

7.1 共分散構造分析

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。

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、実習を行う。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、名義尺度や順序尺度をモデルに適用することは困難である。そこで、これらの尺度を用いて相関係数を算出するための、多分相関（polychoric correlation）、多分系列相関（polyserial correlation）、四分相関（tetrachoric correlation）を算出方法やその適用例を紹介する。

7.1.1 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, 豊田 2009）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 7.1 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。四角は実際に観測された変数を示し（実測変数），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。

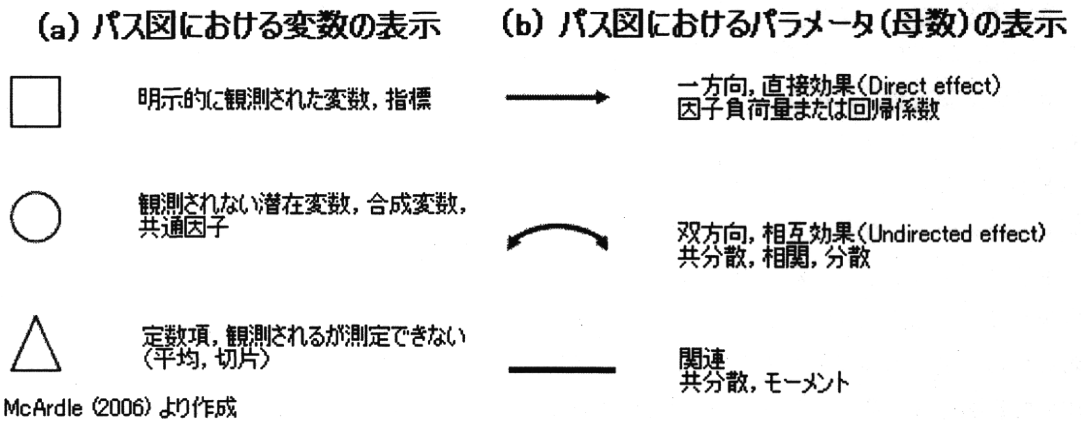


図 7.1 構造方程式をグラフ化（パス図）した場合の表示形式

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 7.2 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。

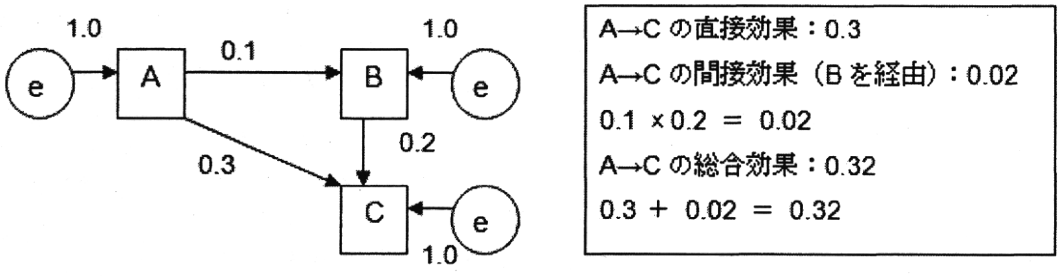


図 7.2 直接効果，間接効果，総合効果の計算例

7.1.2 共分散構造モデルの分析過程

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，近年では R での分析事例を扱ったテキストも多くなり、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) 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。

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-μ], [C-Σ]} 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）。

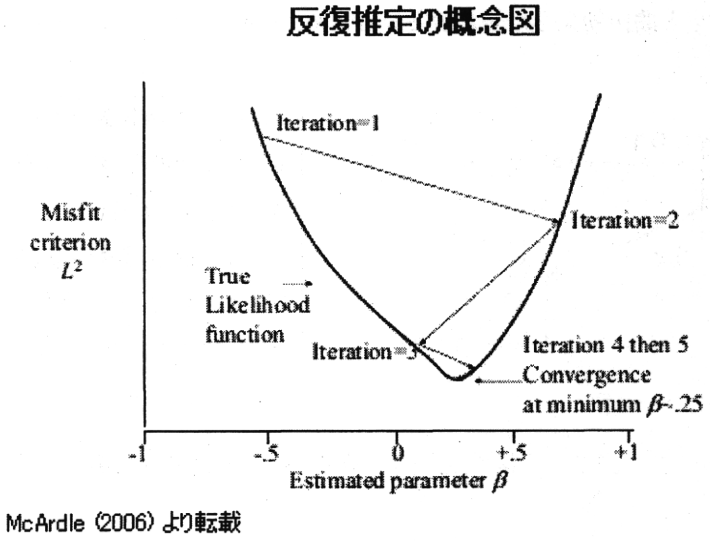


図 7.3 反復推定の概念図

7.2 縦断調査を用いた場合の共分散構造分析モデル

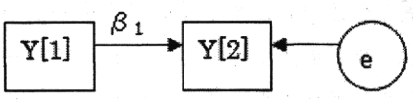
7.2.1 2時点における構造方程式モデル

はじめに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。

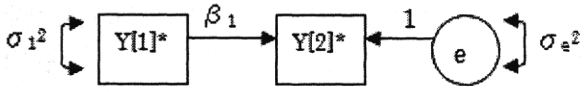
(a) 繰り返し測定するデータを用いた自己回帰モデル

ここでは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。以下は、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。

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



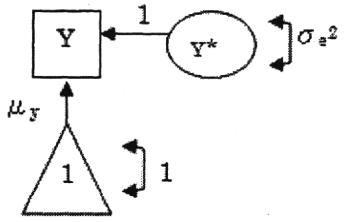
次に共分散を用いた自己回帰モデルを想定すると以下のようになる。アスタリスク（*）は平均周辺の偏差を示す。



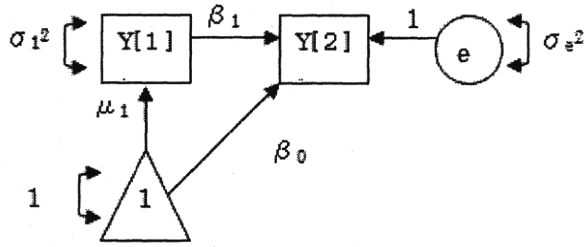
ここで，Y に対する平均と分散は以下のように定義される。μ_0 はグループ内の定数項を示し，Y_n*は各サンプルの平均周辺の偏差（$Y_n - \mu$）を示す。

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

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



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 7.4）。

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

図 7.4 自己回帰モデルによる推定値の解釈

(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

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。

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

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

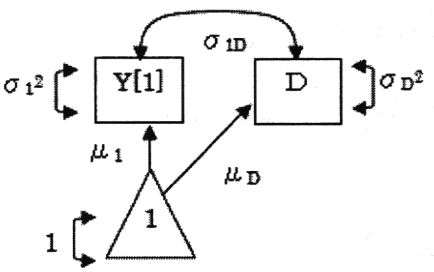
差分の平均と分散，Y[1] との共分散は以下のようになる。

$$D_n = \mu_D + D_n^*$$

$$E\ \{D_n^{*2}\} / (N\text{-}1) = E\ \{D^* D^*\} = \sigma_D^2$$

$$E\ \{Y[1]^* D^*\} = \sigma_{1D}^2$$

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

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。

$$\Delta y_D = \mu_d + \Delta y_D^*$$

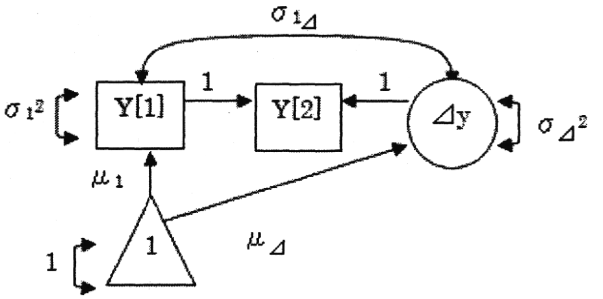
$$E\ \{\Delta y_D^{*2}\} / (N\text{-}1) = E\ \{\Delta y^* \Delta y^*\} = \sigma_\Delta^2$$

$$E\ \{Y[1]^* \Delta y^*\} = \sigma_{1\Delta}^2$$

※ 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。Δy は観測されないため，平均・分散・標準偏差は Y[2]=Y[1]+ Δy 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

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。

7.2.2 不完全なパネルデータを用いた場合の対策

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，補間（Imputation）：データの情報をもとに欠損値（" missing" data）や平均値を修正，補間する方法である。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)

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。

7.2.3 カテゴリカル・データを用いた共分散構造分析

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 7.5）。

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，四分相関係数（Tetrachoric correlation coefficient）である。四分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

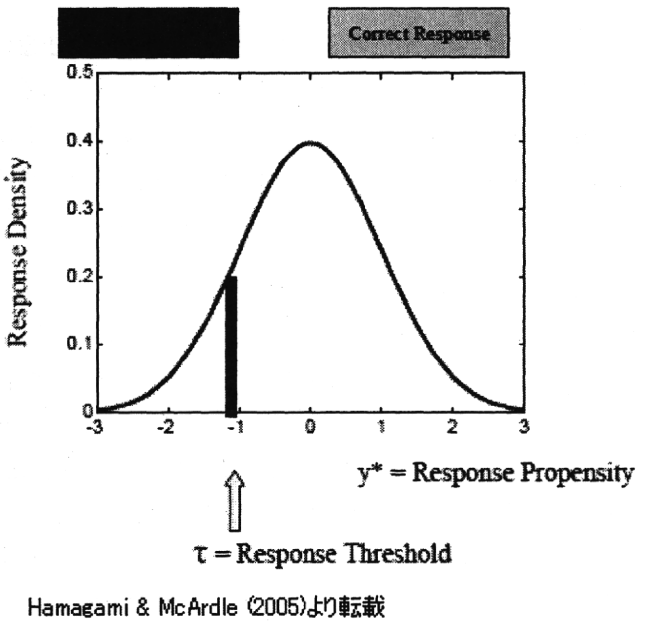
測定される 2 値変数が繰り返し測定される場合を想定し，ロジスティック回帰モデルを適用する場合を取り上げる。ロジスティック回帰モデルは以下のようになる。P(g) / (1-P(g)) はオッズを示す。

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

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，

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

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



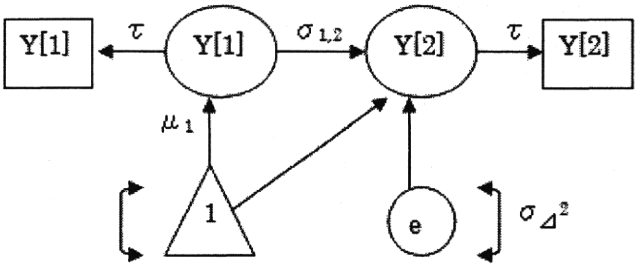
Hamagami & McArdle (2005)より転載

図 7.5 Response propensity の仮定と境界点の関係

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。

また、順序尺度とその他の尺度での相関係数の算出については、多分相関や多分系列相関といったものがあり、多分相関係数は順序尺度同士の相関係数を示し、多分系列相関係数は順初尺度と連続尺度の相関係数を示す。また、二値変数と連続尺度の相関には相関係数の絶対値を用いる相関比を用いる。



7.3 共分散分析の問題点とモデルの適合度評価

共分散構造分析はモデル構築において，変数間の関係を研究者自身が任意に決定することができることが特徴であるが，そのことはモデルの信頼性の低下と表裏一体である。そのため，さまざまなモデルの適合度検定を用いることによってモデルの信頼性を担保することが必要になる。ここでは，竹内・豊田 (1992) や McArdle and Hamagami (2006) に従い共分散構造分析の適合度評価の指標をまとめる。

7.3.1 χ^2 検定

母数の推定方法が最尤推定法または一般化最小二乗法であるとき，「標本数が十分に大きいときに $\chi^2 = (\mathrm{N}-1)\mathrm{f}$ が自由度 df=(n(n+1))/2 − P の χ^2 分布に近似的に従うことを利用して」(p.100)，帰無仮説 (H_0：モデルは正しい) を検定する方法である（f：各推定法の適合度関数，P：母数)。この検定方法の問題点は，サンプル数が大きくなるに連れて検定力が大きくなり，サンプルが大きいために有用なモデルが棄却され，サンプルが小さく「いいかげんな」モデルが棄却されないことが起きうる点にある。$\chi^2 = 0$ であるとき，完全にモデルが適合していると判断される。

7.3.2 適合度指標（GFI：Goodness of Fit Index)

GFI は「構成したモデルが標本共分散行列，あるいは標本相関行列を説明する割合」(p.101) を示した指標であり，決定係数に相当する指標である。χ^2 とは異なりサンプル数に依存しないが，自由度による影響を受ける。よって母数が多くなると GFI は 1 に近づき，モデルの適合度が増加する。

7.3.3 修正適合度指標（AGFI：Adjusted GFI)

AGFI は自由度に依存する GFI から自由度の影響を除した指標である。自由度修正済み決定係数に相当する。

7.3.4 残差平方平均平方根（RMR：Root Mean Square Resident）

RMRは「モデルが説明できなかった標本共分散（標本相関）行列の残量を1セルあたりの平均として示す指標」(p.102)である。RMRはGFIと同様，母数が多くなると最小値である0に近づくという性質がある。指標として解釈が難しく，「各変数の分散が1に基準化されている標本相関行列を用いた場合」(p.102)にのみ明確に解釈ができる。

7.3.5 情報量基準（Information Criterion）：AIC

情報量基準は，「統計モデルの説明力と安定性を統合した指標として平均対数尤度，期待平均対数尤度という2つの測度を導入し，期待平均対数尤度の推定量を共分散構造モデルの評価基準」(p.103)とした指標である。

情報量基準についてはさまざまな指標が存在し，AIC（Akaike's Information Criterion）が一般的である。AICは「最大対数尤度からモデルの自由母数の数を引いた統計量が期待平均対数尤度の近似的な不偏推定量」(p.103)となることを利用した指標であり，AICが最小のモデルが最も適合度が高いとされる。値に絶対的な意味がないため，モデル選択のための比較基準として利用される。

7.3.6 平均二乗誤差（RMSEA：Root Mean Squared Error of Approximation）

RMSEAはサンプル数，自由度の影響をともに考慮した指標であり，$(\chi^2/\text{df-1})/(\text{N-1})$の平方根によって示される。その値が0.05以下であるときにモデルが良いと判断され，0.1以上であると不適合であると判断される。共分散構造分析の適合度として最も一般的な指標であるといえる。

McArdle and Hamagami (2006)においては，χ^2検定によるモデル検定とRMSEAによるモデルの適合度の2つを用いてモデル評価することを推奨している。

7.4 Rによる分析例—自己回帰モデルによる共分散構造分析—

ここでは分析例として、21世紀成年者縦断調査の第1回調査（2002年調査）と第4回調査（2005年調査）を用いて、独身者の結婚意欲の変化を就業形態に着目した分析を行う。また、社会調査データにて一般的な名義尺度や順序尺度を用いて共分散構造分析を行う際の相関係数の算出についての例を提示する。

使用するデータは21世紀成年者縦断調査の履歴データである。成年者調査は結婚や移動による脱落が多く生じており（守泉・釜野2009, 釜野2010)、分析に使用するサンプル数を十分に確保する見地から第4回までのデータを使用した。

結婚意欲の変化と就業形態の変化の関係は「21世紀成年者縦断調査」の結果報告（第4回）において指摘されており（厚生労働省大臣官房統計情報部2006)、その間に就業状態が変化したケースのうち、非正規就業から正規就業へ移行した場合、結婚意欲が高くなるという結果が得られてい

る。経済変動の影響が結婚・出生行動に影響を及ぼすように、個人の景気見通しが結婚意欲に影響を与え、景気見通しに悲観的であると結婚見込みが低くなる（第一生命経済研究所 2008）。女性も正規・非正規で結婚意欲が異なることが指摘されている（永瀬 2002, 福田 2006）。年収の効果をみても同様の傾向である。その他の結婚意欲に与える影響としては、年齢は 30 代から顕著に減少する。学歴が高くなると結婚意欲が高くなる（小林 2006；福田 2006）。このように、若者の就業形態の非正規就業化は、結婚・出生行動への先行指標である結婚意欲に大きく影響する要素であるといえる。

分析に使用する統計ソフトは R である。R はフリーソフトであり、計算機能からグラフィカルな図表の作成まで、様々な分析が可能である。R はオープンソースであることから、新たな関数群（以下「パッケージ」と呼ぶ）が開発・更新されており、最新の分析手法のためのパッケージが様々な研究者によって生み出されている。本分析に主として用いるパッケージは、共分散構造分析を行うことができる sem パッケージ（ver. 0.9-21）、多分相関が算出できる polycor（ver. 0.7-8）、心理学に関する関数が多く格納され、四分相関が算出できる psych（ver. 1.0-92）である。

R において特別の分析手法を適用するときには、パッケージを読み込む必要がある。

パッケージのインストール

```
library(sem)
library(polycor)
library(psych)
```

次に、データを男女別に分ける。ここでは、semd というデータを予め読み込んでおき、そこから性別（Sex）が男性（1）のみを抽出した semdm データと女性（0）を抽出した semdf データを作成している。

サブデータセットを男女別で作成

```
semdm <- semd[semd$Sex == 1,]
semdf <- semd[semd$Sex == 0,]
```

男女別のデータセットから、本分析で使用する変数を、その変数の尺度を定義付けて抽出する。抽出する変数は、第 1 回結婚意欲（w1mrmt：順序尺度）、第 4 回結婚意欲（w4mrmt：順序尺度）、第 4 回時点年齢（w4age：比例尺度）、第 4 回時点学歴（w4edu：順序尺度）、第 4 回時点所得（対数化）（w4inc：比例尺度）、第 1 回時点親との同居の有無（w1cp：名義尺度）、第 1 回から第 4 回までに非正規就業から正規就業への変化の有無（wcemp：名義尺度）の 7 変数である。以下は女性モデルを例に説明する。semdf データから当該変数をそれぞれデータフレームとして呼び出し、尺度の定義を行っている（as.〜）。そのデータを semdf2 とした。summary() は、離散変数であれば、それぞれのカテゴリに含まれる度数と欠損値を、連続変数であれば、最小値、25 パーセンタイル、中央値、平均値、75 パーセンタイル、最大値、欠損値を示す。

サブデータセットの尺度を定義し、記述統計を参照

```
semdf2 <- data.frame(w1mrmt=as.ordered(semdf$W1MrMt),
    w4mrmt=as.ordered(semdf$W4MrMt),
    w4age=as.numeric(semdf$W4Age),
    w4edu=as.ordered(semdf$W4Edu),
    w4inc=as.numeric(semdf$W4LInc),
    w1cp=as.factor(semdf$W1CP),
    wcemp=as.factor(semdf$WCEmp))
summary(semdf2)
```

例）summary アウトプット

```
> summary(semdf2)
  w1mrmt       w4mrmt          w4age
 1   : 111   1   : 103   Min.   :23.00
 2   : 311   2   : 239   1st Qu.:25.00
 3   :1008   3   : 774   Median :28.00
 4   :1600   4   :1401   Mean   :28.64
 5   :1454   5   :1201   3rd Qu.:32.00
NA's: 167   NA's: 933   Max.   :37.00
                        NA's   :81.00

 w4edu          w4inc
1   :  46   Min.   :  0.000
2   :1161   1st Qu.:  4.970
3   : 983   Median :  5.384
4   :1071   Mean   :  5.240
5   :1089   3rd Qu.:  5.704
6   :  77   Max.   :  8.143
NA's: 224   NA's   :806.000

  w1cp        wcemp
0   : 928   0   :2564
1   :3715   1   : 335
NA's:   8   NA's:1752
```

このようにデータセットが出来たら、まずは共分散構造分析に必要な相関係数行列を算出する。Rのsemパッケージでは相関行列または共分散行列をデータセットとして用いて構造モデルを推定する。作成したデータセットは、名義尺度、順序尺度、比例尺度が含まれているため、ピアソンの積率相関係数はもちろん算出してはならない。そこで、カテゴリ変数を用いる際の相関係数である多分相関や多分系列相関、四分相関を算出して、相関行列を作成する。

polycorパッケージにはhetcor（heterogeneous correlation matrix）という関数があり、定義した尺度に応じて相関係数を算出する。hetcor( ) 関数の引数は、hetcor(データ,ML=T（最尤推定（T)、2段階推定（デフォルト)) ,std.err=T（標準誤差を算出する（T)、算出しない（デフォルト))) が基本となり、その他にもいくつかのオプションがある。

use=c("pairwise.complete.obs")) はそれぞれ対応する変数間の欠損値は除くという意味で("complete.obs") と設定すると、全ての変数の数値がそろったデータのみで算出することになり、欠損値が多くある変数が1つでも含まれると算出するのが困難になる。

相関係数の算出

```
hetcf <- hetcor(semdf2,ML=T,std.error=T,use=c("pairwise.complete.obs"))
hetcf
```

例) hetcor( ) アウトプット

```
> hetcf <- hetcor(semdf2,ML=T,std.error=T,use=c("pairwise.complete.obs"))
> hetcf

Maximum-Likelihood Estimates

Correlations/Type of Correlation:
         w1mrmt     w4mrmt      w4age      w4edu      w4inc       w1cp      wcemp
w1mrmt        1 Polychoric Polyserial Polychoric Polyserial Polychoric Polychoric
w4mrmt   0.6648          1 Polyserial Polychoric Polyserial Polychoric Polychoric
w4age    -0.159    -0.2307          1 Polyserial    Pearson Polyserial Polyserial
w4edu   0.09088     0.1403    -0.1201          1 Polyserial Polychoric Polychoric
w4inc     0.072     0.1004     0.1472     0.1089          1 Polyserial Polyserial
w1cp    0.09536      0.119     -0.105    0.07489    -0.06109          1 Polychoric
wcemp   0.02131      0.109      -0.46     0.3167     -0.1545     0.0154          1
```

算出されたアウトプットをみると、下三角行列に相関係数、上三角行列に算出した相関係数のタイプが示される。さらに、オプションに従って標準誤差と相関係数に対応する度数と二値変数の正規性検定のp値行列が算出される。

```
Standard Errors/Numbers of Observations:
         w1mrmt  w4mrmt   w4age   w4edu   w4inc    w1cp wcemp
w1mrmt     4484    3619    4404    4269    3741    4478  2829
w4mrmt  0.01139    3718    3718    3601    3237    3712  2541
w4age   0.01557 0.01635    4570    4427    3845    4562  2899
w4edu   0.01714 0.01847 0.01557    4427    3733    4420  2809
w4inc   0.01742 0.01853 0.01578 0.01707    3845    3837  2668
w1cp    0.02293 0.02501  0.0205 0.02238 0.02404    4643  2894
wcemp   0.03335 0.03457 0.02713 0.02999 0.02907 0.04466  2899


P-values for Tests of Bivariate Normality:
          w1mrmt     w4mrmt     w4age     w4edu      w4inc w1cp
w1mrmt
w4mrmt 2.821e-36
w4age  9.881e-15 2.704e-28
w4edu  1.133e-11  0.002212 1.051e-69
w4inc  4.932e-99 2.042e-85 4.61e-141 1.81e-156
w1cp       0.0819   0.07218 5.749e-21 2.551e-06 4.853e-105
wcemp      0.6625     0.913 4.335e-48 5.266e-11  1.308e-28 <NA>
```

要約で得られる係数は小数第3〜5位の大雑把なものであるため、詳細を知りたいときには以下のように指定するとよい。hetcor関数を算出したhetcfデータの係数 (correlations) と相関係数のタイプ (type) を算出するという指令である。

相関係数行列ならびに係数のタイプの行列の出力

```
hetcf$correlations
hetcf$type
```

例）hetcor( ) 関数から得られた係数とタイプを個別に出力したアウトプット

```
> hetcf$correlations
            w1mrmt     w4mrmt      w4age       w4edu       w4inc
w1mrmt  1.00000000  0.6648295 -0.1589642  0.09087967  0.07200355
w4mrmt  0.66482947  1.0000000 -0.2307165  0.14032880  0.10042899
w4age  -0.15896416 -0.2307165  1.0000000 -0.12011541  0.14719188
w4edu   0.09087967  0.1403288 -0.1201154  1.00000000  0.10893140
w4inc   0.07200355  0.1004290  0.1471919  0.10893140  1.00000000
w1cp    0.09536370  0.1189556 -0.1049946  0.07488511 -0.06108521
wcemp   0.02131452  0.1089523 -0.4600324  0.31665087 -0.15454942

                w1cp       wcemp
 0.09536370  0.02131452
 0.11895559  0.10895225
-0.10499464 -0.46003241
 0.07488511  0.31665087
-0.06108521 -0.15454942
 1.00000000  0.01539576
 0.01539576  1.00000000

> hetcf$type
     [,1]          [,2]          [,3]          [,4]          [,5]
[1,] ""            "Polychoric"  "Polyserial"  "Polychoric"  "Polyserial"
[2,] "Polychoric"  ""            "Polyserial"  "Polychoric"  "Polyserial"
[3,] "Polyserial"  "Polyserial"  ""            "Polyserial"  "Pearson"
[4,] "Polychoric"  "Polychoric"  "Polyserial"  ""            "Polyserial"
[5,] "Polyserial"  "Polyserial"  "Pearson"     "Polyserial"  ""
[6,] "Polychoric"  "Polychoric"  "Polyserial"  "Polychoric"  "Polyserial"
[7,] "Polychoric"  "Polychoric"  "Polyserial"  "Polychoric"  "Polyserial"

     [,6]          [,7]
"Polychoric"  "Polychoric"
"Polychoric"  "Polychoric"
"Polyserial"  "Polyserial"
"Polychoric"  "Polychoric"
"Polyserial"  "Polyserial"
""            "Polychoric"
"Polychoric"  ""
```

ここで相関行列が完成したわけではない。データセットには2つのダミー変数があり、ダミー変数同士は四分相関を用いる必要がある。四分相関を算出するためには psych パッケージの tetrachoric( ) 関数を用いる。tetrachoric 関数の引数の特定にはいくつかの方法があるが、ここでは単純にデータ自体を引数として設定する方法をとる。そのために、新たなデータセットとして第1回調査時親と同居変数（w1cp）と第1回から第4回までに非正規就業から正規就業への変化の

有無（wcemp）のみのサブデータセットを作成し、四分相関を算出する。最後に四分相関の係数（ロー ρ）と二値変数の閾値（タウ τ）を算出する。

2値変数のみのサブデータセットを作成し、四分相関係数を出力

```
semdf3 <- data.frame(semdf$W1CP,semdf$WCEmp)
summary(semdf3)

tetraf <- tetrachoric(semdf3,correct=T)
tetraf

tetraf$rho
tetraf$tau
```

例）四分相関アウトプット

```
> tetraf <- tetrachoric(semdf3,correct=T)
> tetraf
Call: tetrachoric(x = semdf3, correct = T)
tetrachoric correlation
           semdf.W1CP semdf.WCEmp
semdf.W1CP      1.000       0.015
semdf.WCEmp     0.015       1.000

 with tau of
 semdf.W1CP semdf.WCEmp
      -0.84        1.20


> tetraf$rho
            semdf.W1CP semdf.WCEmp
semdf.W1CP  1.00000000  0.01539576
semdf.WCEmp 0.01539576  1.00000000
> tetraf$tau
 semdf.W1CP semdf.WCEmp
 -0.8420829   1.1974937
```

最後に二値変数と連続尺度の相関係数の絶対値をとり相関比として、相関行列の完成である。

ここから共分散構造分析の分析方法について説明する。第一に相関行列を読み込ませる。read.moments関数で、下三角行列で記述した相関行列を記述する。

```
cormf <- read.moments(diag=T,names=
 c("w1mrmt","w4mrmt","w4age","w4edu","w4inc","w1cp","wcemp"))


 1.00000
 0.66483  1.00000
-0.15896 -0.23072 1.00000
 0.09088  0.14033  -0.12012 1.00000
 0.07200  0.10043  0.14719  0.10893 1.00000
 0.09536  0.11896  0.10499  0.07489 0.06109 1.00000
 0.02131  0.10895  0.46003  0.31665 0.15455 0.01540 1.00000
```

例）相関行列の読み込み済みを確認

```
> cormf
         w1mrmt   w4mrmt    w4age   w4edu   w4inc   w1cp wcemp
w1mrmt  1.00000  0.00000  0.00000 0.00000 0.00000 0.0000     0
w4mrmt  0.66483  1.00000  0.00000 0.00000 0.00000 0.0000     0
w4age  -0.15896 -0.23072  1.00000 0.00000 0.00000 0.0000     0
w4edu   0.09088  0.14033 -0.12012 1.00000 0.00000 0.0000     0
w4inc   0.07200  0.10043  0.14719 0.10893 1.00000 0.0000     0
w1cp    0.09536  0.11896  0.10499 0.07489 0.06109 1.0000     0
wcemp   0.02131  0.10895  0.46003 0.31665 0.15455 0.0154     1
```

共分散構造分析で対象となる相関行列の読み込みを確認し、構造方程式の作成を行う。specify.model( ) 関数によって、モデルを指定する。独立変数Aから独立変数Bへ影響があるとき、「A ->B,b01,NA」というように指定する。b01は推定する母数を示し、NAの部分には固定母数を指定する。NAは固定母数を推定するという意味である。複雑なモデルを推定する場合には、識別問題を解消するために固定母数を1にする等の操作を行うことがある。今回のモデルは単純なものであるため、全ての固定母数の推定が可能である。また、相関を示す場合は「A <−> B」、分散は「A <−> A」となる。e01〜e07は各変数の分散を示している。

モデルの指定

```
modelf <- specify.model()
w1mrmt -> w4mrmt,b01,NA
w1cp -> w1mrmt,b02,NA
wcemp -> w4mrmt,b03,NA
w4edu -> wcemp,b04,NA
w4edu -> w4inc,b05,NA
w4inc -> w4mrmt,b06,NA
w4inc -> wcemp,b07,NA
w4age -> w4mrmt,b08,NA
w1mrmt <-> w1mrmt,e01,NA
w4mrmt <-> w4mrmt,e02,NA
wcemp <-> wcemp,e03,NA
w4age <-> w4age,e04,NA
w4edu <-> w4edu,e05,NA
w4inc <-> w4inc,e06,NA
w1cp <-> w1cp,e07,NA
```

モデルが設定できたら、いよいよ sem( ) でモデルを解く。sem( ) 関数の引き数は、モデル、相関（共分散）行列、サンプル数となる。また、モデルの要約推定量として std.coef( ) は標準化推定値を算出し、mod.indices( ) は修正指数を示す。

モデルの要約

```
semsolf <- sem(modelf,cormf,N=4651)
summary(semsolf)
std.coef(semsolf)
mod.indices(semsolf)
```

例）モデルの解の要約・標準化推定値・修正指数

```
> summary(semsoit)

 Model Chisquare =  2070.6   Df =  13 Pr(>Chisq) = 0
 Chisquare (null model) =  5885.4   Df =  21
 Goodness-of-fit index =  0.90428
 Adjusted goodness-of-fit index =  0.79384
 RMSEA index =  0.18450   90% CI: (0.17784, 0.19124)
 Bentler-Bonnett NFI =  0.64817
 Tucker-Lewis NNFI =  0.43321
 Bentler CFI =  0.64913
 SRMR =  0.10810
 BIC =  1960.9

 Normalized Residuals
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
-10.800   0.000   0.744   2.400   4.910  31.400


 Normalized Residuals
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
-10.800   0.000   0.744   2.400   4.910  31.400

 Parameter Estimates
    Estimate  Std Error z value  Pr(>|z|)
b01  0.619980 0.010690   57.9960 0.0000e+00 w4mrmt <--- w1mrmt
b02  0.095360 0.014598    6.5325 6.4700e-11 w1mrmt <--- w1cp
b03  0.191965 0.011888   16.1478 0.0000e+00 w4mrmt <--- wcemp
b04  0.303415 0.013879   21.8607 0.0000e+00 wcemp <--- w4edu
b05  0.108930 0.014577    7.4725 7.8604e-14 w4inc <--- w4edu
b06  0.059872 0.010661    5.6163 1.9511e-08 w4mrmt <--- w4inc
b07  0.121499 0.013879    8.7539 0.0000e+00 wcemp <--- w4inc
b08 -0.229290 0.012053  -19.0232 0.0000e+00 w4mrmt <--- w4age
e01  0.990906 0.020555   48.2085 0.0000e+00 w1mrmt <--> w1mrmt
e02  0.507989 0.010539   48.1993 0.0000e+00 w4mrmt <--> w4mrmt
e03  0.885146 0.018361   48.2074 0.0000e+00 wcemp <--> wcemp
e04  1.000000 0.020743   48.2086 0.0000e+00 w4age <--> w4age
e05  1.000000 0.020743   48.2086 0.0000e+00 w4edu <--> w4edu
e06  0.988134 0.020497   48.2085 0.0000e+00 w4inc <--> w4inc
e07  1.000000 0.020743   48.2086 0.0000e+00 w1cp <--> w1cp

 Iterations =  0
```

```
> std.coef(semsolf)
        Std. Estimate
b01 b01    0.62344178 w4mrmt <--- w1mrmt
b02 b02    0.09536000   w1mrmt <--- w1cp
b03 b03    0.19303714  w4mrmt <--- wcemp
b04 b04    0.30341511   wcemp <--- w4edu
b05 b05    0.10893000   w4inc <--- w4edu
b06 b06    0.06020668  w4mrmt <--- w4inc
b07 b07    0.12149899   wcemp <--- w4inc
b08 b08   -0.23057051  w4mrmt <--- w4age
e01 e01    0.99090647 w1mrmt <--> w1mrmt
e02 e02    0.51367701 w4mrmt <--> w4mrmt
e03 e03    0.88514593   wcemp <--> wcemp
e04 e04    1.00000000   w4age <--> w4age
e05 e05    1.00000000   w4edu <--> w4edu
e06 e06    0.98813426   w4inc <--> w4inc
e07 e07    1.00000000     w1cp <--> w1cp


> mod.indices(semsolf)

 5 largest modification indices, A matrix:
 wcemp:w4age  w4age:wcemp w1mrmt:w4age  w4inc:w4age w4age:w1mrmt
   1203.2902     984.0683     133.9828     120.8834     117.4975

  5 largest modification indices, P matrix:
 wcemp:w4age w4age:w1mrmt  w4inc:w4age  w4edu:w4age  w1cp:w4mrmt
  1203.29018    133.98278    120.88343     67.09399     54.69364
```

このように、指定したモデルに従ってモデルの推定がなされる。この構造を図として示したのが図7.6である。図の作成には、フリーソフトの「Graphviz」を用いた。Graphvizは DOT 言語を用いてグラフ表現を行い GIF 等のファイルに変換するシステムである。Graphvizのコードと共分散構造分析モデルの R コードは同様のシステムであるため、ある程度そのまま使用することができ、構造モデルを作成するのに使い勝手がよい。例で示した Graphviz のコードは、digraph という図表作成の開始を指示し、sem という名前の図表を作成することを示している。その具体的な設定は { } 内にそれぞれの要素ごとに指示を行う。rankdir=LR によって、上から下方向の縦書きの図表から左から右の方向性をもった横書きの図表となる。四角（観測変数）や丸（潜在変数）で示されるボックスを node といい、node[ ] はその形状やラベルのフォントを規定することができる。矢印は edge であり、同様に様々な形状を規定することができる。[rank=same(min,max)] は node の位置する場所を決めることができ、same は規定された node が同じ水準に並ぶことを指定している。min はこの図の場合は最も左、max は最も右にくるように指定している。ただし構造モデルの設定の仕方によって、望ましい場所に位置しないこともある。graph[ ] では、グラフ全体の設定やラベルをつけることができる。以下は変数ごとの設定がなされ、node と edge との関係に推定値（標準推定値）をラベルとして設定すると、図7.6のような図を作成することができる。