共分散構造分析の紹介

佐藤 健一（原医研・計量生物）

1　はじめに

変数間に仮定されたパス図のもとで共分散を求め, 標本共分散に近づけるようにパス係数を推定する統計手法として共分散構造分析を紹介する.

共分散構造分析は構造方程式モデリング (SEM; Structural Equation Modeling) とよばれることもあり, 従来の重回帰モデル, 因子分析, パス解析モデルを含むとともに, 新たに, これらを組み合わせた多重指標モデルや経時測定データに対する潜在成長曲線モデルなど非常に幅広い統計モデルを含んでいる. ここでは, 観測変数を対象としたパス解析モデルを例に, 数理統計学的な側面を紹介するとともに, フリーの統計ソフトウェア “R”のスクリプトを使った解析例を示す. なお, 解析例には “sem”パッケージを利用した. 解析結果のパス図はフリーウェア Graphviz(AT&T 社, http://www.graphviz.org/) で解釈される dot 言語で記述されており, テキストファイルとして比較的容易に編集することができる.

2　統計モデル

パス解析モデルを念頭に共分散構造モデルを概観する. 統計モデルを理解することで, 解析で仮定されている条件や統計量の意味について理解が深まり, 統計ソフトウェアの誤用を防ぐとともに, 解析結果の解釈も容易となる.

外生変数を $\boldsymbol{x}$, 内生変数を $\boldsymbol{y}$, 内生変数に関する観測誤差を $\boldsymbol{e}$ とすると, パス解析モデルは, 回帰係数行列 B と Γ を用いて次のようにかける.

$$\boldsymbol{y} = B\boldsymbol{y} + \Gamma\boldsymbol{x} + \boldsymbol{e}$$

外生変数と内生変数をまとめて, $\boldsymbol{z} = (\boldsymbol{y}', \boldsymbol{x}')'$ とすると

$$\boldsymbol{z} = R\boldsymbol{z} + \boldsymbol{\varepsilon}$$

ここで,

$$R = \begin{pmatrix} B & \Gamma \\ O & O \end{pmatrix}, \boldsymbol{\varepsilon} = (\boldsymbol{e}', \boldsymbol{x}')', \mathrm{Var}(\boldsymbol{\varepsilon}) = \begin{pmatrix} \mathrm{Var}(\boldsymbol{e}) & O \\ O & \mathrm{Var}(\boldsymbol{x}) \end{pmatrix} = \Omega.$$

したがって, $\boldsymbol{z} = (I - R)^{-1}\boldsymbol{\varepsilon}$, とかける.

変数 $\boldsymbol{z}$ のうち, $G = (I, O)$ を用いて観測変数を $\boldsymbol{w} = G\boldsymbol{z}$ とあらわせば, 未知パラメータ $\boldsymbol{\theta}$ を持つモデルのもとでの共分散 $\Sigma(\boldsymbol{\theta})$ は次のようにかける.

$$\mathrm{Var}(\boldsymbol{w}) = G(I-R)^{-1}\Omega\{(I-R)^{-1}\}'G' = \Sigma(\boldsymbol{\theta})$$

ここでは, パス解析モデルを念頭に R と Ω を与えたが, $\boldsymbol{z}$ に潜在変数が加わっても, R と Ω を与えることができればモデルのもとでの観測変数の共分散を同様に求めることができる.

未知パラメータに関する対数尤度関数は観測変数の標本共分散行列 S と観測変数の個数 $\dim(S)$ を用いて,

$$f(\boldsymbol{\theta}) = -\mathrm{tr}\Sigma(\boldsymbol{\theta})^{-1}S + \log|\Sigma(\boldsymbol{\theta})^{-1}S| + \dim(S)$$

とかけるので, これを最大化することで. $\boldsymbol{\theta}$ の最尤推定値が得られる. なお, $\Sigma(\boldsymbol{\theta}) = S$ という帰無仮説のもとで, $\chi^2 = (N-1)f$ は自由度 df のカイ二乗分布にしたがう. ただし,

$$df = \#S - \#R - \#\Omega,$$

ここで, $\#S = \dim(S) * (\dim(S)+1)/2$, $\#R$ は R で与えたパス係数の個数, $\#\Omega$ は Ω で与えた共分散の個数である.

そして, モデルの適合度は, 例えば, GFI (Goodness of Fit Index) を使って評価でき, 1 に近いほど良いモデルと言える.

$$GFI = 1 - \mathrm{tr}\{\Sigma(\boldsymbol{\theta})^{-1}S - I\}^2/\mathrm{tr}\{\Sigma(\boldsymbol{\theta})^{-1}S\}^2,$$

ここで, $A^2 = AA'$ とする. また, RMSEA (Root Mean Square Error of Approximation) は 0 に近いほど, 良いモデルとなり,

$$RMSEA = \sqrt{(\chi^2/df - 1)/(N-1)}.$$

3 解析例

3.1 パス解析モデル

パス解析モデルは観測変数のみを対象としたパス図に対するモデルで, 従来の重回帰モデルの説明変数をさらに目的変数として階層的な重回帰モデルを構成する. したがって, 最終的にすべての矢印が向かう先は 1 つの目的変数となる. ここでは, 出生体重データ (佐和 隆光: 回帰分析, 朝倉書店,1979) を用いて, パス図と統計モデルとの対応を確認する.

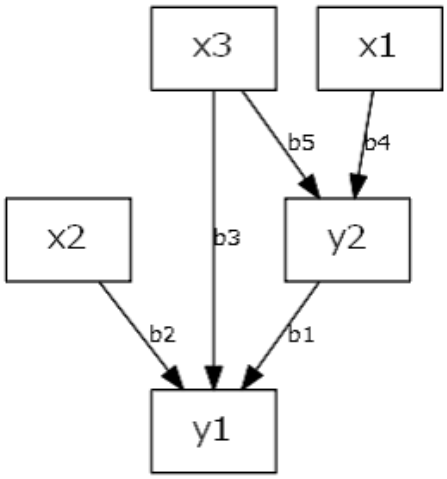

Figure 1: 変数名によるパス図. 四角で囲んだ変数は観測変数を示す. 矢印が刺さる変数を内生変数, 刺さらない変数を外生変数とよぶ. 内生変数には誤差をあらわす確率変数を加えることが多い.

3.1.1　変数名と行列による記述

ここでは, Figure 1 のパス図を持つパス解析モデルを行列で表現する. 外生変数 y_1, y_2 と内生変数 x_1, x_2, x_3 を用いて,

$$\begin{pmatrix} y_1 \\ y_2 \end{pmatrix} = \begin{pmatrix} & b_1 \\ & \end{pmatrix} \begin{pmatrix} y_1 \\ y_2 \end{pmatrix} + \begin{pmatrix} & b_2 & b_3 \\ b_4 & & b_5 \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \\ x_3 \end{pmatrix} + \begin{pmatrix} e_1 \\ e_2 \end{pmatrix},$$

とかける. ここで, $\mathrm{Var}(\boldsymbol{e}) = diag(b_6, b_7)$. したがって, $\boldsymbol{\theta} = (b_1. \cdots, b_7)'$. あるいは, 外生変数と内生変数をまとめて以下のようにかける.

$$\begin{pmatrix} y_1 \\ y_2 \\ x_1 \\ x_2 \\ x_3 \end{pmatrix} = \begin{pmatrix} & & b_1 & & b_2 & b_3 \\ & & & b_4 & & b_5 \\ & & & & & \\ & & & & & \\ & & & & & \end{pmatrix} \begin{pmatrix} y_1 \\ y_2 \\ x_1 \\ x_2 \\ x_3 \end{pmatrix} + \begin{pmatrix} e_1 \\ e_2 \\ x_1 \\ x_2 \\ x_3 \end{pmatrix}$$

$$\mathrm{Var} \begin{pmatrix} e_1 \\ e_2 \\ x_1 \\ x_2 \\ x_3 \end{pmatrix} = \begin{pmatrix} b_6 & & & & \\ & b_7 & & & \\ & & S_{3,3} & S_{3,4} & S_{3,5} \\ & & S_{4,3} & S_{4,4} & S_{4,5} \\ & & S_{5,3} & S_{5,4} & S_{5,5} \end{pmatrix}$$

ここで, モデルの自由度は, $df = 5(5+1)/2 - 5 - \{2 + 3(3+1)/2\} = 2$ と計算できる.

3.1.2 DOT 言語によるパス図の記述

DOT 言語の最小の表現は, 例えば, x から y に矢印を引くパス図をかく場合,

```
digraph{x -> y}
```

となる. ファイルの拡張子は, “*.gv”あるいは “*.dot”などが使われる. ファイル内で日本語を記述する場合にはファイルのエンコードを UTF-8(BOM なし) として保存する必要がある. パス図の全体を知らなくても, 2 つの要素について記述していくだけで, 結果的に最適な配置が得られる, 以下に, Figure 1 のパス図を記述した DOT 言語を示す.

[DOT 言語のスクリプト]

```
digraph "res" {
  rankdir=TB;
  size="8,8";
  node [fontname="meiryo" fontsize=14 shape=box];
  edge [fontname="meiryo" fontsize=10];
  center=1;
  "y2" -> "y1" [label="b1"];
  "x2" -> "y1" [label="b2"];
  "x3" -> "y1" [label="b3"];
  "x1" -> "y2" [label="b4"];
  "x3" -> "y2" [label="b5"];
}
```

3.1.3 パス図とスクリプト

[R のスクリプト]

```
N <- 15
S <- matrix(c(
  0.147, 0,0,0,0,
  2.141, 47.924, 0,0,0,
  0.059, 11.919, 22.695, 0,0,
  0.714, -23.443, 2.900, 236.171,0,
 -0.049, 0.029, 0.629, 0.086,0.257),
  5, 5, byrow=TRUE)
rownames(S) <- colnames(S) <- c(
  "BabyWeight","MamWeight","MamAge","PregnancyDays","Smoking")

S <- S+t(S)
diag(S) <- diag(S)/2

library(sem)
model <- specifyEquations()
BabyWeight = b1*MamWeight+b2*PregnancyDays+b3*Smoking
MamWeight  = b4*MamAge+b5*Smoking
V(BabyWeight) = e1
V(MamWeight) = e2

res <- sem(model, S, N=N,standardized=F,
  fixed.x=c("MamAge","PregnancyDays","Smoking"))

pathDiagram(res,file="Result",graphics.fmt="png",edge.labels="both",
  node.font=c("meiryo", 14),edge.font=c("meiryo", 10))
```

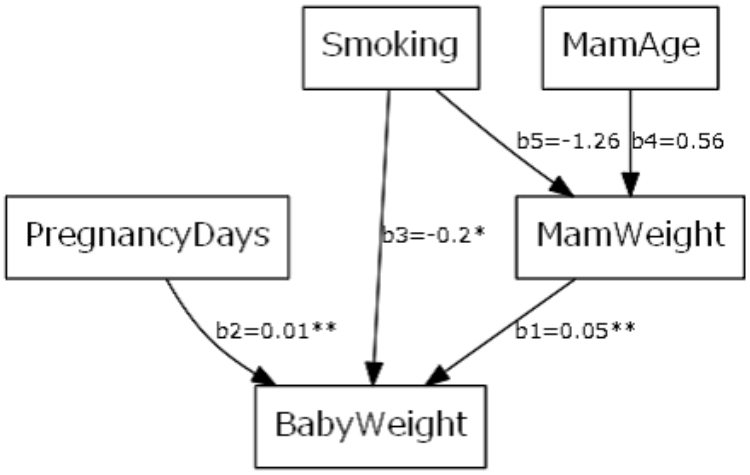

Figure 2: パス解析モデルの解析結果のパス図. 出生体重に対する母親の年齢, 喫煙, 体重, 妊娠日数の影響を示す. ただし, $GFI = 0.875$ なのでモデルの当てはまりはあまり良くない. 解釈としては, 例えば, 母親の年齢が 10 才増えると母親の体重が 5.6kg 増え, 母親の体重が 10kg 増えると出生体重が 0.5kg 増える, となる.

[R によるスクリプトの実行結果 (抜粋)]

```
Model Chisquare =  7.132198   Df =  2 Pr(>Chisq) = 0.02826591
 Goodness-of-fit index =  0.8572732
 Adjusted goodness-of-fit index =  -0.07045068
 RMSEA index =  0.4281271   90% CI: (NA, NA)
 Bentler CFI =  0.7838957
 AIC =  21.1322

   Estimate Std Error z value Pr(>|z|)
b1    0.049     0.006   7.617    0.000     BabyWeight <--- MamWeight
b2    0.008     0.003   2.754    0.006 BabyWeight <--- PregnancyDays
b3   -0.199     0.087  -2.278    0.023       BabyWeight <--- Smoking
b4    0.560     0.373   1.500    0.134         MamWeight <--- MamAge
b5   -1.258     3.508  -0.359    0.720        MamWeight <--- Smoking
e1    0.027     0.010   2.646    0.008    BabyWeight <--> BabyWeight
e2   41.285    15.604   2.646    0.008      MamWeight <--> MamWeight
```

3.2 多重指標モデル

多重指標モデルはパス解析モデルのように観測変数だけでなく, 因子分析の潜在変数も対象としてパス図を構成できる. ここでは, 食物摂取とがんの関係データ (豊田秀樹, 前田忠彦, 柳井 晴夫: 原因をさぐる統計学 - 共分散構造分析入門 (ブルーバックス), 講談社, 1992) を用いる.

3.2.1 変数名と行列による記述

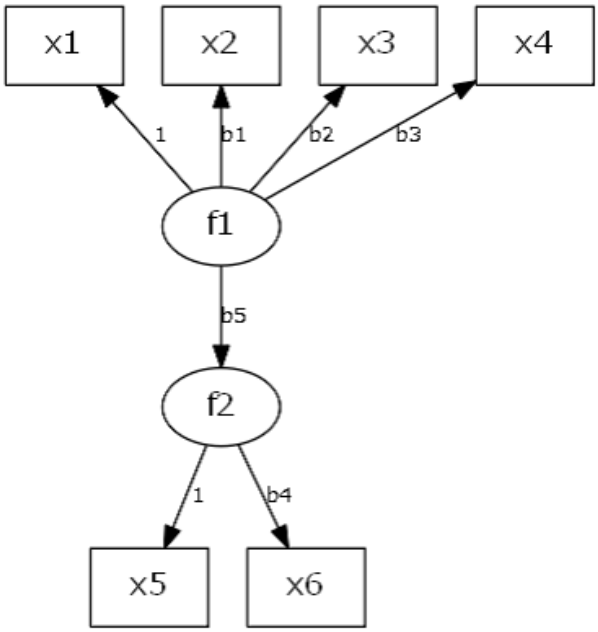

Figure 3: 変数名によるパス図. 四角で囲んだ変数は実際に観測される観測変数, 丸で囲んだ変数は観測されない潜在変数を示す.

Figure 3 には実際には観測されることのない潜在変数がある. 潜在変数は因子分析においては因子として利用されている. ここでは, 行列表現における潜在変数の扱いに注意したい. 観測変数 $x1, \cdots, x_6$, 潜在変数 f_1, f_2 を用いて,

$$\begin{pmatrix} x_1 \\ x_2 \\ x_3 \\ x_4 \\ x_5 \\ x_6 \\ f_1 \\ f_2 \end{pmatrix} = \begin{pmatrix} & 1 & \\ & b_1 & \\ & b_2 & \\ & b_3 & \\ & & 1 \\ & & b_4 \\ & & \\ & b_5 & \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \\ x_3 \\ x_4 \\ x_5 \\ x_6 \\ f_1 \\ f_2 \end{pmatrix} + \begin{pmatrix} e_1 \\ e_2 \\ e_3 \\ e_4 \\ e_5 \\ e_6 \\ f_1 \\ e_7 \end{pmatrix}$$

$$\mathrm{Var}(e_1, \cdots, e_6, f_1, e_7) = \mathrm{diag}(b_6, \cdots, b_{11}, b_{12}, b_{13})$$

とかける. なお, S との比較のために観測変数のみの共分散行列を求めるときは $G = (I_6, O_{2\times 2})$ を用いて, $\Sigma(\boldsymbol{\theta}) = G(I - B)\Omega(I - B)t(G)$ とする.

3.2.2 パス図とスクリプト

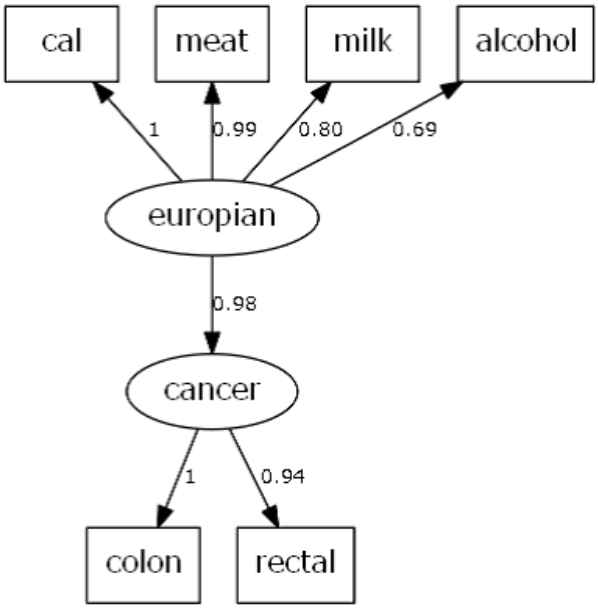

Figure 4: 多重指標モデルの解析結果のパス図. 洋食傾向 (総熱量, 肉類, 乳製品, 酒類) が下部消化管のガン傾向 (大腸ガン, 直腸ガン) に与える影響を示す. ここで, 洋食傾向と下部消化管のガン傾向はともに潜在変数であり, この多重指標モデルは 2 つの因子分析の潜在変数に対する単回帰に対応している.

[R のスクリプト]

```
N <- 47
S <- matrix(c(
  1,0,0,0,0,0,
  0.773,1,0,0,0,0,
  0.716,0.723,1,0,0,0,
  0.634,0.452,0.232,1,0,0,
  0.739,0.853,0.579,0.588,1,0,
  0.810,0.672,0.479,0.610,0.752,1),
  6, 6, byrow=TRUE)

S <- S+t(S)
diag(S) <- diag(S)/2
rownames(S)<- colnames(S)<- c("cal","meat",
  "milk","alcohol","colon","rectal")

library(sem)
model.cfa <- cfa(reference.indicators=T)
europian: cal, meat, milk, alcohol
cancer: colon, rectal
```

```
model.reg <- specifyEquations()
cancer = b1(0.98)*europian
V(cancer) = e1

model<- combineModels(model.cfa, model.reg)
res <- sem(model, S, N,  objective=objectiveML)
```

3.3　複雑なモデル

多重指標モデルの中でも潜在変数に対する回帰構造が若干, 複雑なモデルの適用例として, Wheaton et al. (Sociological methodology, 1977) のデータを扱う. なお, 解析例は Fox (Structural Equation Modeling, 2006) にも紹介されている.

3.3.1　パス図とスクリプト

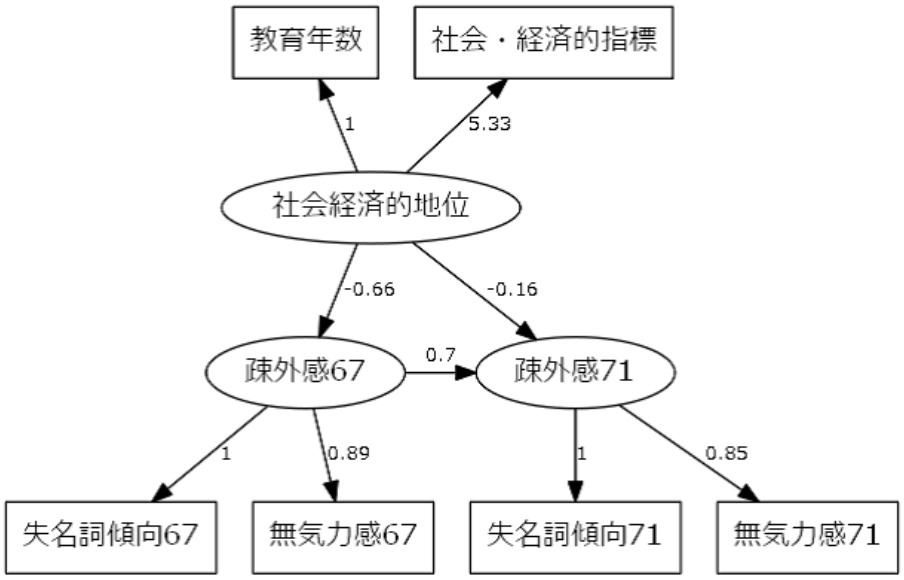

Figure 5: 複雑なモデルの解析結果のパス図

[R のスクリプト]

```
N=932
S <- matrix(c(
  11.834, 0, 0, 0, 0, 0,
  6.947, 9.364, 0, 0, 0, 0,
  6.819, 5.091, 12.532, 0, 0, 0,
  4.783, 5.028, 7.495, 9.986, 0, 0,
  -3.839, -3.889, -3.841, -3.625, 9.610, 0,
```

```
  -21.899, -18.831, -21.748, -18.775, 35.522, 450.288),
  6, 6, byrow=TRUE)
rownames(S) <- colnames(S) <- c("失名詞傾向 67","無気力感 67",
  "失名詞傾向 71","無気力感 71","教育年数","社会・経済的指標")

library(sem)
model.cfa <- cfa(reference.indicators=T)
疎外感 67: 失名詞傾向 67, 無気力感 67
疎外感 71: 失名詞傾向 71, 無気力感 71
社会経済的地位: 教育年数, 社会・経済的指標

model.reg <- specifyEquations()
疎外感 71 = b1*疎外感 67
疎外感 67 = b2*社会経済的地位
疎外感 71 = b3*社会経済的地位

model<- combineModels(model.cfa, model.reg)
res <- sem(model, S, N)
```

3.4 潜在成長曲線モデル

潜在成長曲線モデルは平均構造を持つモデルであり, 実際には共分散行列だけでなく, 観測値の平均も利用している. 例として, Potthoff and Roy (Biometrika,1964) に掲載されている歯科矯正に関する経時測定データを用いる. このデータに対する回帰分析については,

藤井良宜, 佐藤健一, 冨田哲治, 和泉志津恵
医療系のための統計入門 (事例でわかる統計シリーズ), 実教出版, 2015,

にも解説がある.

3.4.1 パス図とスクリプト

["Potthoff.csv"]

```
Age8,Age10,Age12,Age14,Male,Const
26  , 25  , 29  , 31  ,   1,    1
21.5, 22.5, 23  , 26.5,   1,    1
23  , 22.5, 24  , 27.5,   1,    1
25.5, 27.5, 26.5, 27  ,   1,    1
20  , 23.5, 22.5, 26  ,   1,    1
24.5, 25.5, 27  , 28.5,   1,    1
22  , 22  , 24.5, 26.5,   1,    1
24  , 21.5, 24.5, 25.5,   1,    1
23  , 20.5, 31  , 26  ,   1,    1
27.5, 28  , 31  , 31.5,   1,    1
23  , 23  , 23.5, 25  ,   1,    1
21.5, 23.5, 24  , 28  ,   1,    1
17  , 24.5, 26  , 29.5,   1,    1
22.5, 25.5, 25.5, 26  ,   1,    1
23  , 24.5, 26  , 30  ,   1,    1
22  , 21.5, 23.5, 25  ,   1,    1
21  , 20  , 21.5, 23  ,   0,    1
21  , 21.5, 24  , 25.5,   0,    1
20.5, 24  , 24.5, 26  ,   0,    1
```

```
23.5, 24.5, 25  , 26.5,   0,    1
21.5, 23  , 22.5, 23.5,   0,    1
20  , 21  , 21  , 22.5,   0,    1
21.5, 22.5, 23  , 25  ,   0,    1
23  , 23  , 23.5, 24  ,   0,    1
20  , 21  , 22  , 21.5,   0,    1
16.5, 19  , 19  , 19.5,   0,    1
24.5, 25  , 28  , 28  ,   0,    1
```

[R のスクリプト]

```
d <- read.csv("Potthoff.csv")

library(sem)
model <- specifyModel()
i -> Age8, NA, 1
i -> Age10, NA, 1
i -> Age12, NA, 1
i -> Age14, NA, 1
s -> Age8, NA, -3
s -> Age10, NA, -1
s -> Age12, NA, 1
s -> Age14, NA, 3
i <-> i, vi,1
s <-> s, vs,1
i <-> s, vis,1
Age8 <-> Age8, vy,2
Age10 <-> Age10, vy,2
Age12 <-> Age12, vy,2
Age14 <-> Age14, vy,2
Const -> i, b1i,17
Const -> s, b1s,0.5
Male -> i, bmi,-1
Male -> s, bms,0.3

res <- sem(model, S=rawMoments(d), N=nrow(d),
  raw=T, fixed.x=c("Const","Male"))
```

統計相談について

統計についてお困りのことがあれば, ホームページをご覧の上, お問い合わせ下さい.

http://home.hiroshima-u.ac.jp/ksatoh/

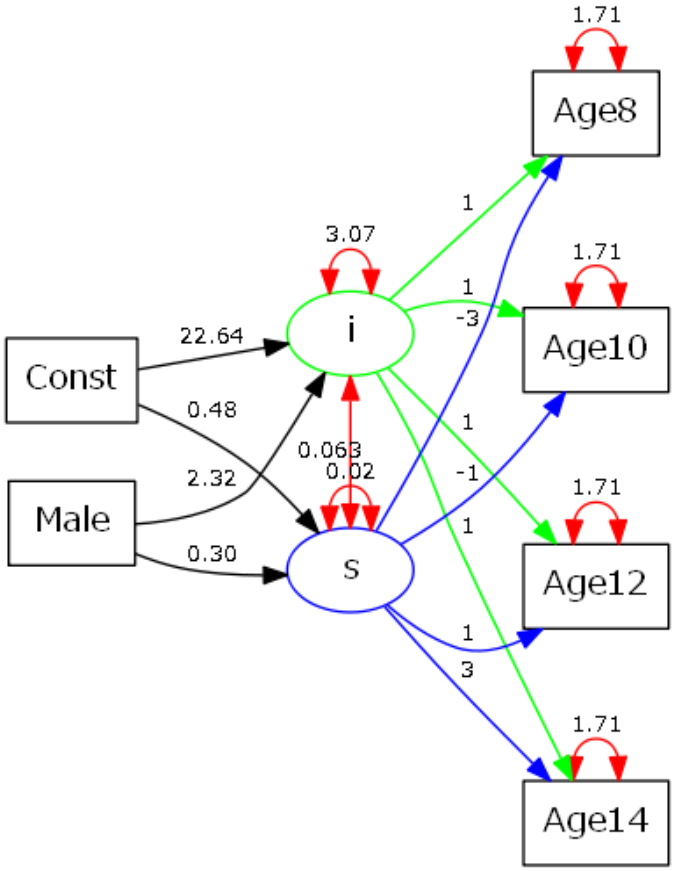

Figure 6: 潜在成長曲線モデルの解析結果のパス図. 赤色は分散・共分散を, 緑色はパス係数を 1 に固定した切片 (i) からのパスを, 青色はパス係数を-3,-1,1,3 と固定した傾き (s) からのパスを示す. 女性および男性に対する成長曲線が, それぞれ, $y = 22.64 + 0.48x$ および $y = (22.64 + 2.32) + (0.48 + 0.30)x$ であることがわかる.