mwp-142

SEM による CFA へのアプローチ

確証的因子分析 (CFA: confirmatory factor analysis) においては直接観測することのできない変数 — **潜在変数** (latent variables) — を想定してモデルが記述され、分析が行われます。そのための有力な手段となるものが**構造方程式モデリング** (SEM: structural equation modeling) による推定法であるわけですが、本 whitepaper ではシミュレーションデータを用いてその有用性を検証してみます。

1. 想定モデル

今、100 人の受験生を対象にした数学の試験結果に関するデータが与えられたとします。変数 y_1, y_2, y_3 は模擬試験の点数を、変数 y_4 は入学試験での点数を表すものとした上で、次のようなモデルを想定することにします。

$$\begin{aligned} y_1 &= \alpha_1 + \beta_1 X + \varepsilon_1 \\ y_2 &= \alpha_2 + \beta_2 X + \varepsilon_2 \\ y_3 &= \alpha_3 + \beta_3 X + \varepsilon_3 \\ y_4 &= \alpha_4 + \beta_4 X + \varepsilon_4 \end{aligned} \tag{1}$$

ここでポイントとなるのはそれぞれの学生の数学的資質を表すものとした潜在変数 X の存在です。X の値自体は観測できるわけではありませんが、観測された変数値 y_1, y_2, y_3, y_4 はその影響を反映したものであるとするのがモデル式 (1) の本質であるわけです。

2. シミュレーションデータの生成

ここでは正規分布に従う擬似乱数発生用の関数 `rnormal(`m, s`)` を用いて y_1, y_2, y_3, y_4 の値を生成することにします。ただし m は平均値を、s は標準偏差の大きさを意味します[*1]。以下に示したコードでは

$$\alpha_1 = \alpha_2 = \alpha_3 = 50,\ \alpha_4 = 100$$
$$\beta_1 = \beta_2 = \beta_3 = 1,\ \beta_4 = 2$$

というパラメータ設定をしていますが、特に深い意味があるわけではありません。また m, s の設定値も X 用、ε 用とで共通のものとしていますが、同一である必要性は全くありません。なお、コード上、X ではなく X_0 という変数名を用いている点に注意してください。後述する `sem` コマンド中で

```
. sem (y1 y2 y3 y4 <- X)
```

という記述が出てくるわけですが、そこで用いる潜在変数名 X との混同を避けるため、ここでは X_0 という名称を用いています。X_0 はシミュレーションデータの生成過程で用いられるだけであり、y_1, y_2, y_3, y_4 生成後はデータセットから削除してしまっても構いません。実際、X_0 は潜在変数という位置付けなので、データセット中には本来存在しないものです。後段では `sem` による予測値 $\widehat{X}$ と X_0 との対比なども行いますが、それはシミュレーションだからこそできる操作だという点に注意してください。

```
. set seed 11
. set obs 100
. generate X0 = round(rnormal(0,10))
. generate y1 = round(50 + X0 + rnormal(0, 10))
. generate y2 = round(50 + X0 + rnormal(0, 10))
. generate y3 = round(50 + X0 + rnormal(0, 10))
. generate y4 = round(100 + 2*X0 + rnormal(0, 10))
```

このコードを実行することにより 100 個のデータが生成されるわけですが、参考までに先頭 10 個のデータをリスト出力しておくと次のようになります。

```
. list in 1/10
```

	X0	y1	y2	y3	y4
1.	9	66	42	66	122
2.	-3	60	58	49	107
3.	0	58	51	54	90
4.	-20	38	34	23	54
5.	-9	35	32	55	68

[*1] `rnormal()` や `round()` といった関数の仕様については `[FN]` **Random-number functions** (*mwp-076*) をご参照ください。

6.	28	85	61	80	157
7.	0	48	43	46	90
8.	-5	49	42	64	94
9.	16	69	44	58	132
10.	6	36	54	40	111

この 100 個のデータからなるデータセット `cfa01.dta` は

`http://www.math-koubou.jp/sdata14.html`

よりダウンロードすることができます。

3. OLS 推定

この例の場合、模擬試験の成績 — 例えば y_1 — と入学試験での成績 y_4 との間には変数 X の存在により相関が生じます（モデル式 (1) 参照)。実際、y_1 と y_4 について散布図を作成してみると次のようになります。

```
. twoway (scatter y4 y1), yscale(range(0 200)) ylabel(0(50)200)
> xscale(range(0 100)) xlabel(0(20)100) aspectratio(1)
```

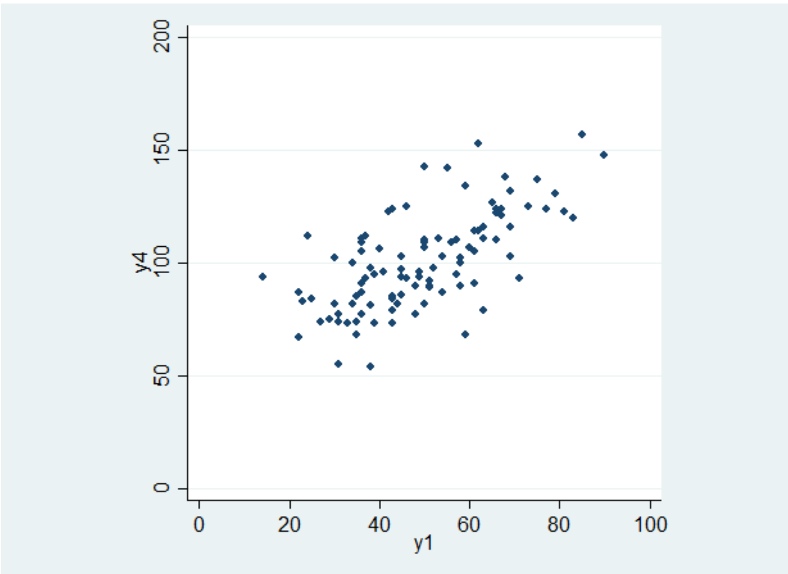

図 1　y_1 - y_4 散布図

このようなグラフから y_4 と y_1 の間で線形回帰を実行してみたくなりますがそれは誤った結果を導くことになります。モデル式 (1) が成り立つとした場合、y_1 は内生変数であり、線形回帰の説明変数として使用した場合には OLS (ordinary least squares) 推定の前提条件に抵触することになるからです。この点を無視してグラフ中に線形回帰直線をプロットさせると次のような図となります。

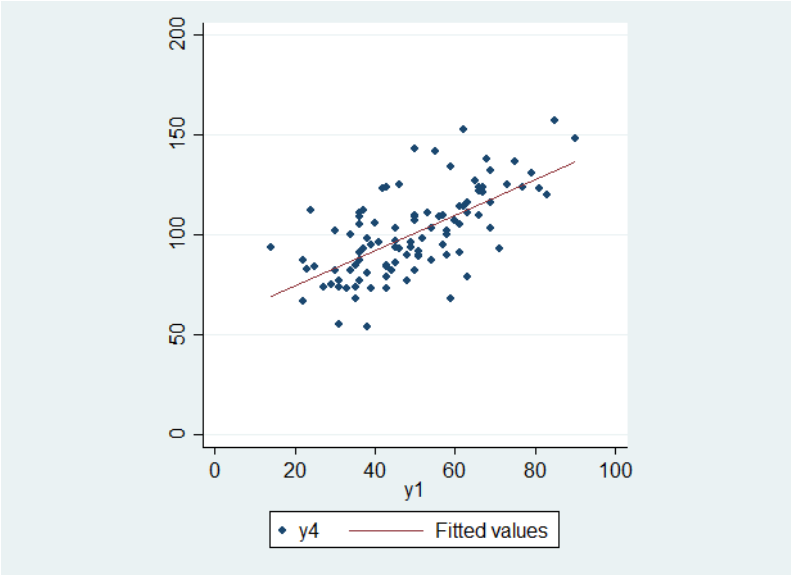

図 2　線形回帰直線のフィット

図 1 を人間の目で見た場合には傾きが 2 に近い直線が期待されるわけですが、`regress` コマンドによってフィットされる直線の傾きは 1 に近いものとなります。

```
. regress y4 y1 *2
```

```
. regress y4 y1

      Source |       SS           df       MS      Number of obs   =       100
-------------+----------------------------------   F(1, 98)        =     72.79
       Model |  19626.3913         1  19626.3913   Prob > F        =    0.0000
    Residual |  26425.3187        98  269.646109   R-squared       =    0.4262
-------------+----------------------------------   Adj R-squared   =    0.4203
       Total |   46051.71        99  465.168788   Root MSE        =    16.421

------------------------------------------------------------------------------
          y4 |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
          y1 |   .8879716   .1040821     8.53   0.000     .6814241    1.094519
       _cons |   56.81541   5.407421    10.51   0.000     46.08456    67.54626
------------------------------------------------------------------------------
```

すなわち

$$\widehat{y_4}^{(1)} = 0.89 \cdot y_1 + 56.82 \tag{2}$$

というのが OLS 推定の回帰式であるわけですが、この式から算出される予測値 $\widehat{y_4}^{(1)}$ を実際の y_4 の値と対向させてプロットしてみると次のようなグラフとなります。

*2 メニュー操作： Statistics ▷ Linear models and related ▷ Linear regression

```
. predict y4hat1, xb [*3]
. twoway (scatter y4hat1 y4) (function x, range(y4)),
> ytitle(y4hat1) yscale(range(40 165)) ylabel(50 100 150)
> xtitle(y4) xscale(range(40 165)) xlabel(50 100 150)
> title(OLS estimation) legend(order(1 "OLS predictions" 2 "perfect fit"))
> aspectratio(1)
```

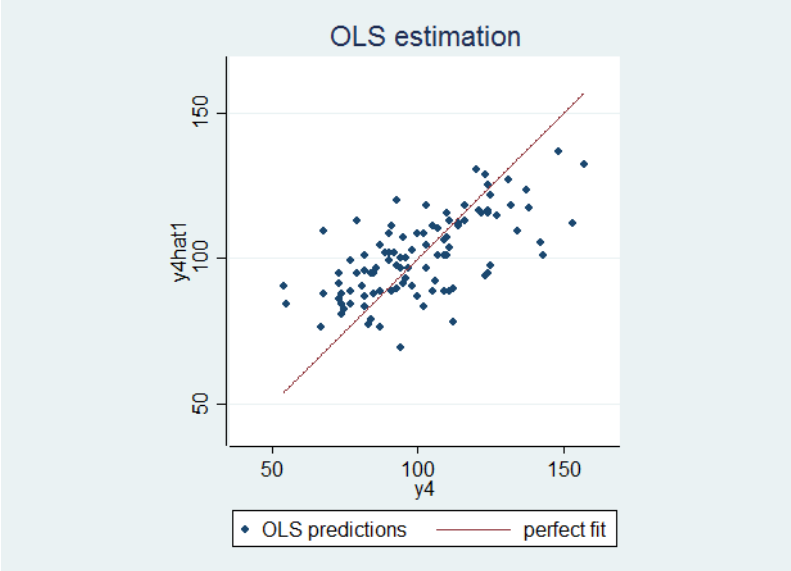

図 3　OLS 推定 – $\widehat{y}_4^{(1)}$

Perfect fit を表す直線 $\widehat{y}_4^{(1)} = y_4$ からはかなり外れた点が数多く存在しており、回帰式 (2) による予測の精度は高いとは言えないことがわかります。

[*3] メニュー操作： Statistics ▷ Postestimation ▷ Predictions ▷ Predictions and their SEs, leverage statistics, distance statistics, etc. ▷ Launch と操作

4. SEM 推定

(1) モデルのフィット

今度はモデル式 (1) を忠実に表現する SEM モデルを SEM Builder 上で構成します[*4]。

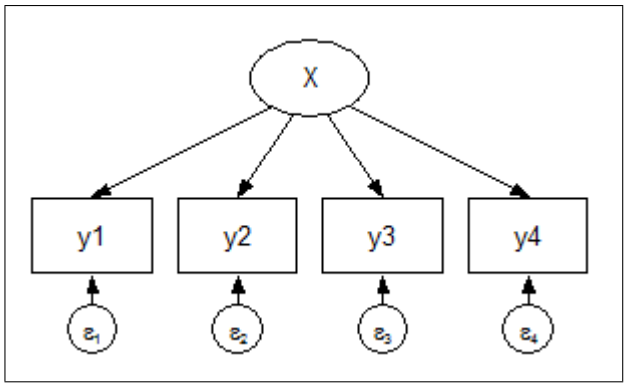

図 4 SEM モデル

このモデルでは潜在変数名を X としていますが、その値はモデルフィットによって推定されるものです。データセット中には X_0 という変数も存在していますが、それとは異なるものである点に注意してください。
 アイコンを使って推定を実行すると次のような出力を得ることができます。

```
. sem (X -> y1, ) (X -> y2, ) (X -> y3, ) (X -> y4, ), latent(X ) nocapslatent

Endogenous variables

Measurement:  y1 y2 y3 y4

Exogenous variables

Latent:       X

Fitting target model:

Iteration 0:   log likelihood = -1602.0669
Iteration 1:   log likelihood = -1601.9957
Iteration 2:   log likelihood = -1601.9954
Iteration 3:   log likelihood = -1601.9954
```

[*4] SEM Builder の操作方法については *mwp-123* 等をご参照ください。

```
Structural equation model                       Number of obs     =        100
Estimation method  = ml
Log likelihood     = -1601.9954

 ( 1)  [y1]X = 1
```

	OIM Coef.	Std. Err.	z	P>|z|	[95% Conf.	Interval]
Measurement						
y1 <-						
X	1	(constrained)				
_cons	49.5	1.577688	31.38	0.000	46.40779	52.59221
y2 <-						
X	.766656	.1295992	5.92	0.000	.5126462	1.020666
_cons	49.32	1.391537	35.44	0.000	46.59264	52.04736
y3 <-						
X	.8901337	.1307325	6.81	0.000	.6339027	1.146365
_cons	50.76	1.441466	35.21	0.000	47.93478	53.58522
y4 <-						
X	1.677966	.2239866	7.49	0.000	1.23896	2.116971
_cons	100.77	2.145966	46.96	0.000	96.56398	104.976
var(e.y1)	115.8384	20.88314			81.35782	164.9323
var(e.y2)	115.4232	18.07179			84.92212	156.8793
var(e.y3)	102.3447	17.82931			72.74094	143.9965
var(e.y4)	85.84477	34.57717			38.98171	189.0457
var(X)	133.0716	33.74149			80.95725	218.7334

```
LR test of model vs. saturated: chi2(2)   =      1.70, Prob > chi2 = 0.4267
```

モデルをコマンドインタフェース上で規定する場合には

```
sem (X -> y1 y2 y3 y4)
```

のように記述します。

表が始まる直前に

```
( 1) [y1]X = 1
```

という表記がありますが、これは y_1 に関する方程式の X に対する係数値、すなわちモデル式 (1) で言う β_1 が 1 と仮定されたことを示しています。これは正規化 (normalization) に伴う要請に基づき sem によって自動設定されたものですが、その背景情報については [SEM] **intro 4** (*mwp-122*) をご参照ください。sem による出力中から推定結果を抽出すると次のようになります。

パラメータ推定値			シミュレーション条件		
$\widehat{\beta}_1$	=	1	β_1	=	1
$\widehat{\beta}_2$	=	0.77	β_2	=	1
$\widehat{\beta}_3$	=	0.89	β_3	=	1
$\widehat{\beta}_4$	=	1.68	β_4	=	2
$\widehat{\alpha}_1$	=	49.5	α_1	=	50
$\widehat{\alpha}_2$	=	49.32	α_2	=	50
$\widehat{\alpha}_3$	=	50.76	α_3	=	50
$\widehat{\alpha}_4$	=	100.77	α_4	=	100

モデル式 (1) 中のパラメータについては良好な推定値が得られていることがわかります。なお、テーブルの後段に示されているのは $\varepsilon_1, \varepsilon_2, \varepsilon_3, \varepsilon_4$ 及び X の分散推定値です。

(2) $\widehat{X}$ の算出

sem による推定を実行した後、predict コマンドを使用すると、潜在変数や観測可能変数の予測値を求めることができます。最初に潜在変数 X の予測値 $\widehat{X}$ を算出してみます。

```
. predict Xhat, latent(X)
```
[*5]


predict コマンドの用法については [SEM] **example 14** (*mwp-140*) をご参照ください。

[*5] メニュー操作： Statistics ▷ SEM (structural equation modeling) ▷ Predictions ▷ Factor scores

この predict コマンドの実行によって $\widehat{X}$ の値を含む新変数 Xhat が生成されたわけですが、その値をシミュレーション時に使用した X_0 と対向させる形でプロットしてみます。

```
. twoway (scatter Xhat X0) (function x, range(X0)),
> ytitle(Xhat) xtitle(X0) title(Predicted values of X)
> legend(order(1 "SEM prediction" 2 "perfect fit")) aspectratio(1)
```

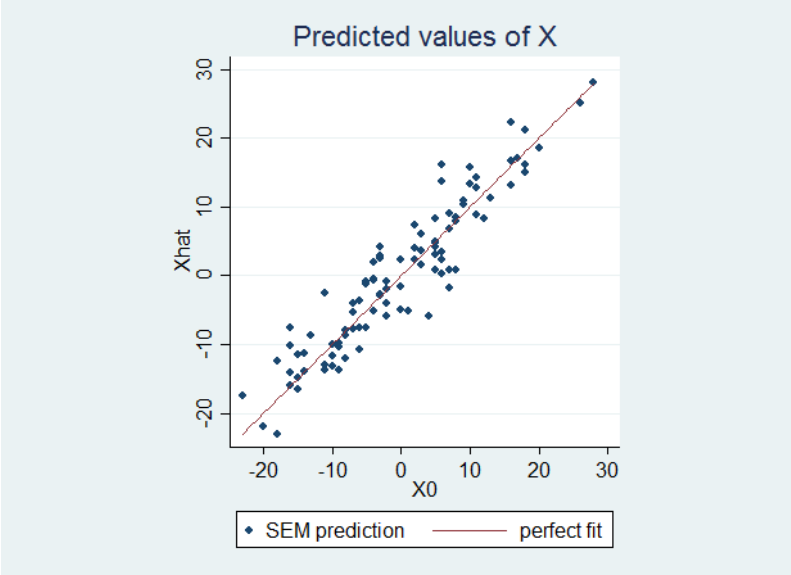

図 5 SEM 推定 – $\widehat{X}$

X_0 と $\widehat{X}$ の値とは比較的良好な対応が取れていることがわかります。通常のデータセットの場合、潜在変数の真の値を知ることはできませんが、この例ではシミュレーションによってデータを生成しているため、このような検証が可能となるわけです。

(3) $\widehat{y_4}^{(2)}$ の算出

セクション 3 では OLS 推定による予測値 $\widehat{y_4}^{(1)}$ を算出したわけですが、ここでもやはり入学試験時の成績 y_4 について SEM 推定による予測値 $\widehat{y_4}^{(2)}$ を算出し、真の値である y_4 との対比を行ってみます。

```
. predict y4hat2, xb(y4)
```
[*6]


predict コマンドの用法については [SEM] **example 14** (*mwp-140*) をご参照ください。

[*6] メニュー操作： Statistics ▷ SEM (structural equation modeling) ▷ Predictions ▷ Observed endogenous variables

この predict コマンドの実行によって $\widehat{y}_4^{(2)}$ の値を含む新変数 y4hat2 が生成されたわけですが、その値を y_4 と対向させる形でプロットしてみます。

```
. twoway (scatter y4hat2 y4) (function x, range(y4)),
> ytitle(y4hat2) yscale(range(40 165)) ylabel(50 100 150)
> xtitle(y4) xscale(range(40 165)) xlabel(50 100 150)
> title(SEM estimation) legend(order(1 "SEM predictions" 2 "perfect fit"))
> aspectratio(1)
```

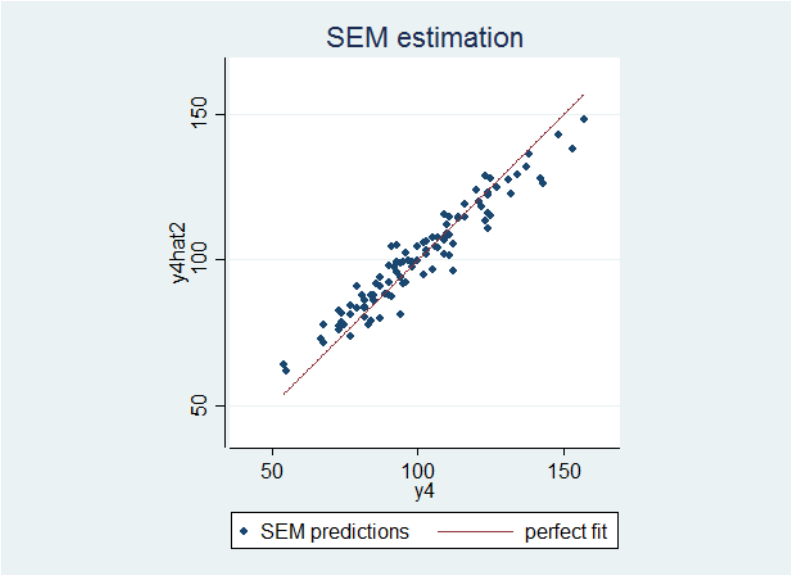

図 6　SEM 推定 – $\widehat{y}_4^{(2)}$

OLS 推定の場合のグラフ（図 3）と比較すると明らかなように、perfect fit を表す直線 $\widehat{y}_4^{(2)} = y_4$ の近傍に予測値が凝集しており、それだけ予測の精度が改善していることがわかります。

データセット中には模擬試験の結果を表す変数が y_1, y_2, y_3 という形で 3 つ含まれています。この数が 2 以下の場合には識別が得られず、sem コマンドの実行に支障を来すことになるので注意する必要があります。

■