

example 27g - 測定モデル（単一因子）【評価版】

本 whitepaper は GSEM を単一因子の測定モデルに適用したときの用例を紹介するものです。

1. 一般化 SEM の特徴
2. データセット
3. GSEM プロビットモデル
4. 混成モデル

1. 一般化 SEM の特徴

gsem は sem ではサポートされていない次の 2 つの機能を提供します。

- 一般化線形応答変数を含む SEM のフィット
- 多階層混合 SEM のフィット

これら 2 つの機能は組み合わせて使用することもできます。

一般化線形応答変数 (generalized linear response variables) とは応答変数が一般化線形モデル (GLM: generalized linear models) によって規定されるものであることを意味します。プロビット/ロジスティック回帰モデル、順序プロビット/ロジスティック回帰モデル、多項ロジスティック回帰モデル等がこれに該当します。

多階層混合モデル (multilevel mixed models) では多階層のグループレベル効果をモデル化することができます。

2. データセット

[SEM] example 27g には Example データセット gsem_1fmm.dta を用いた GSEM の用例が紹介されています。

```
. use http://www.stata-press.com/data/r13/gsem_1fmm.dta *1
(single-factor pass/fail measurement model)
```

このデータセット中には潜在変数 X の測定値を表す変数 x_1, x_2, x_3, x_4 の値が 123 件記録されています。

[SEM] example 1 (mwp-123) でも単一因子測定モデルのフィットを行ったわけですが、そのときの x_1, x_2, x_3, x_4 は連続変数でした。これに対し gsem_1fmm.dta の場合、 x_1, x_2, x_3, x_4 はいずれも合否を表す 2 値変数である点が異なります。これらの変数についての情報は注釈中に記されています。

```
. notes list *2
```

```
. notes list

_dta:
1. Fictional data.
2. The variables x1, x2, x3, and x4 record 1=pass, 0=fail.
3. Pass/fail for x1, x2, x3: score > 100
4. Pass/fail for x4: score > 725
5. Variable s4 contains actual score for test 4.
```

x_1, x_2, x_3 の場合、点数が 101 以上のときに合格、 x_4 の場合には点数が 726 以上のときに合格と判定されることがわかります。なお、 s_4 という変数中にはテスト 4 のときの実際の点数が記録されています。

3. GSEM プロビットモデル

ここでは x_1, x_2, x_3, x_4 が潜在変数 X の顕在化したものとするプロビットモデルを設定します。

- Statistics > SEM (structural equation modeling) > Model building and estimation と操作
- SEM Builder 上で  アイコンをクリックし、gsemへの切替えを行う
-  アイコンを選択し、描画領域内の任意の点をクリック、表示される Measurement component ダイアログ上で次の設定を行う
- Main タブ: Latent variable name: X
 Measurement variables: x1 x2 x3 x4
 Make measurements generalized: Family/Link: Bernoulli, Probit

*1 メニュー操作： File > Example Datasets > Stata 13 manual datasets と操作、Structural Equation Modeling Reference Manual [SEM] の Example 27g の項よりダウンロードする。

*2 メニュー操作： Data > Data utilities > Notes utilities > List or search notes

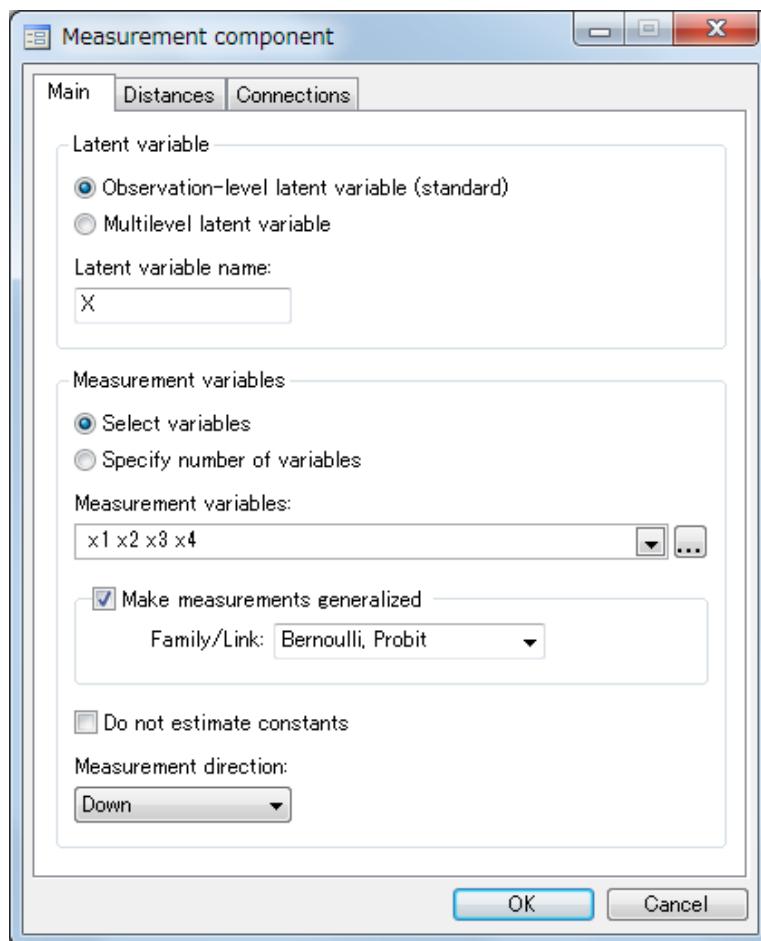


図 1 Measurement component ダイアログ - Main タブ

SEM Builder 上の描画領域内には次のようなモデルが構成されます。

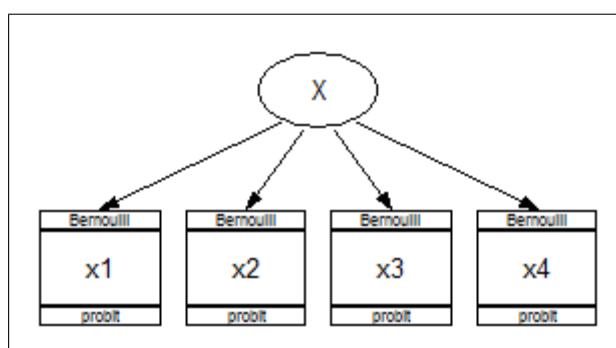


図 2 GSEM プロビットモデル

図 2 に示された測定モデルを数式で表現すると次のようになります。

$$\begin{aligned}\Pr(x_1 = 1) &= \Phi(\alpha_1 + X\beta_1) \\ \Pr(x_2 = 1) &= \Phi(\alpha_2 + X\beta_2) \\ \Pr(x_3 = 1) &= \Phi(\alpha_3 + X\beta_3) \\ \Pr(x_4 = 1) &= \Phi(\alpha_4 + X\beta_4)\end{aligned}\tag{1}$$

ただし $\Phi(\cdot)$ は $N(0, 1)$ 正規累積分布関数を表します。

今、通常の連続値を用いてスコアリングされるテストについて考えてみましょう。スコアを s_1, s_2, s_3, s_4 とし、それぞれ $N(\mu_i, \sigma_i^2)$ という正規分布に従うものとします。さらに各々のテストに対し c_i というカットオフ値を設け、 $s_i \geq c_i$ の場合に合格として扱うことになります。テストスコアが与えられたとするなら、

$$s_i = \gamma_i + X\delta_i + \epsilon_i, \quad \epsilon_i \sim N(0, \sigma_i^2)\tag{2}$$

という線形回帰モデルを設定することになるでしょう。しかし gsem_1fmm.dta というデータセット中には s_i の値は記録されておらず（ただし s_4 を除く）合否の結果を示す x_i の情報しか記録されていないわけです。

そこで合否判定の問題について考察することにしましょう。ある学生がテスト i に合格する確率は学生の点数がカットオフ値を上回る確率によって規定されることになります。すなわち

$$\begin{aligned}\Pr(s_i > c_i) &= \Pr(\gamma_i + X\delta_i + \epsilon_i > c_i) \\ &= \Pr\{\epsilon_i > c_i - (\gamma_i + X\delta_i)\} \\ &= \Pr\{-\epsilon_i \leq -c_i + (\gamma_i + X\delta_i)\} \\ &= \Pr\{-\epsilon_i \leq (\gamma_i - c_i) + X\delta_i\} \\ &= \Pr\{-\epsilon_i/\sigma_i \leq (\gamma_i - c_i)/\sigma_i + X\delta_i/\sigma_i\} \\ &= \Phi\{(\gamma_i - c_i)/\sigma_i + X\delta_i/\sigma_i\}\end{aligned}$$

のように変形できるわけですが、最後の式はプロピットモデルに他なりません。

(1) 式と対比させるなら

$$\begin{aligned}\alpha_i &= (\gamma_i - c_i)/\sigma_i \\ \beta_i &= \delta_i/\sigma_i\end{aligned}\tag{3}$$

という関係があることがわかります。(3) 式の右辺は共に σ_i — 線形モデルにおける誤差項の標準偏差 — で割られている点に注意してください。合否に関するデータの場合には元々のスコアによって表されるスケール情報は失われます。傾き β_i について言えば、推定可能なのは線形モデルにおける傾き δ_i を σ_i で割った値であるわけです。切片 α_i についても同様で、推定できるのは $(\gamma_i - c_i)$ という開きの値を σ_i で割ったスケールを持たない値となります。

Builder 上で  アイコンをクリックすると次のような GSEM estimation options ダイアログが表示されます
が、デフォルトの設定のままモデルを実行します。

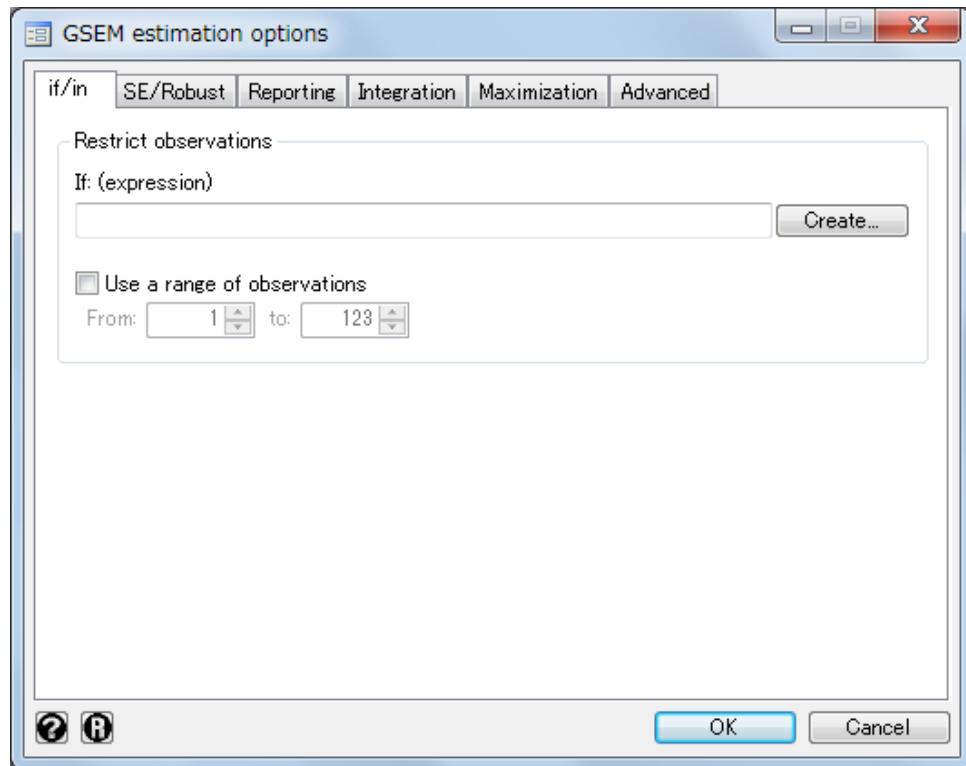


図 3 GSEM estimation options ダイアログ

```
. gsem (X -> x1, family(bernoulli) link(probit)) (X -> x2, family(bernoulli) li-
> nk(probit)) (X -> x3, family(bernoulli) link(probit)) (X -> x4, family(bernou-
> lli) link(probit)), latent(X ) nocapslatent

Fitting fixed-effects model:

Iteration 0: log likelihood = -329.82091
Iteration 1: log likelihood = -329.57665
Iteration 2: log likelihood = -329.57664

Refining starting values:

Grid node 0: log likelihood = -273.75437

Fitting full model:

Iteration 0: log likelihood = -273.75437
Iteration 1: log likelihood = -264.3035
Iteration 2: log likelihood = -263.37815
Iteration 3: log likelihood = -262.305
Iteration 4: log likelihood = -261.69025
Iteration 5: log likelihood = -261.42132
Iteration 6: log likelihood = -261.35508
Iteration 7: log likelihood = -261.3224
```

Iteration 8: log likelihood = -261.3133						
Iteration 9: log likelihood = -261.30783						
Iteration 10: log likelihood = -261.30535						
Iteration 11: log likelihood = -261.30405						
Iteration 12: log likelihood = -261.30337						
Iteration 13: log likelihood = -261.30302						
Iteration 14: log likelihood = -261.30283						
Iteration 15: log likelihood = -261.30272						
Iteration 16: log likelihood = -261.30267						
Iteration 17: log likelihood = -261.30264						
Iteration 18: log likelihood = -261.30263						
Generalized structural equation model				Number of obs =		123
Log likelihood = -261.30263						
(1) [x1]X = 1						
	Coef.	Std. Err.	z	P> z	[95% Conf. Interval]	
x1 <- X _cons	1 (constrained) -.3666763	.1896773	-1.93	0.053	-.738437	.0050844
x2 <- X _cons	1.33293 -.4470271	.4686743 .2372344	2.84 -1.88	0.004 0.060	.4143455 -.911998	2.251515 .0179438
x3 <- X _cons	.6040478 -.2276709	.1908343 .1439342	3.17 -1.58	0.002 0.114	.2300195 -.5097767	.9780761 .0544349
x4 <- X _cons	9.453342 -4.801027	5.151819 2.518038	1.83 -1.91	0.067 0.057	-.6440375 -9.736291	19.55072 .1342372
var(X)	2.173451	1.044885			.847101	5.576536

なお、コマンドインターフェースの場合には次のような指定で実行が行えます。

```
. gsem (x1 x2 x3 x4 <-X), probit
```

なお、probit という指定は family(bernoulli) link(probit) という指定と同義です。実行結果に関しては次の点に注意してください。

1. X は潜在外生変数であるため、正規化のための制約 (normalizing constraint) が必要となります。この例では X は最初の観測可能変数 x_1 に anchoring され、そのパス係数に対して 1 という制約が課されています。正規化については [SEM] intro 4 (mwp-122) をご参照ください。
2. $X \rightarrow x_2$, $X \rightarrow x_3$ のパス係数はそれぞれ 1.33, 0.60 と推定されています。一方、 $X \rightarrow x_4$ のパス係数推定値は 9.45 となっています。

4. 混成モデル

評価版では割愛しています。

