LightStone

Stata で簡単に試せる SEM

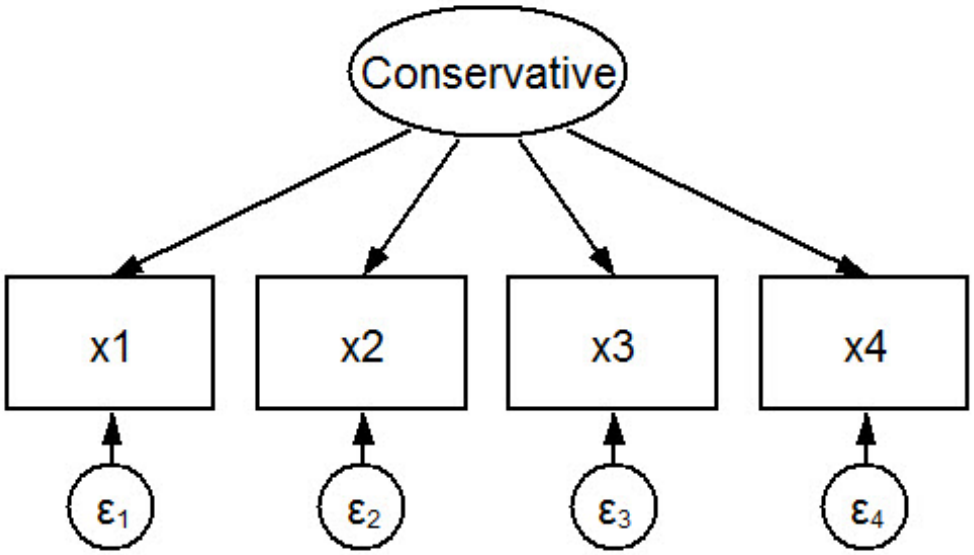
第2回　モデル推定の詳細

SEM の第二回目です. 今回は初回の内容の細部を少し詳しく考察することにします. 解説には主に Stata のマニュアル [SEM] STATA STRUCTURALEQUATION MODELING REFERENCE MANUAL を用います.

第1章　推定の実際

1.1　モデル推定

- 第一回と同じく nlsy97cfa.dta を用いて極めてシンプルなモデルを推定します.
- ここではモデル推定の舞台裏について解説することを主目的します.
- 例として次に示す 4 つの因子からなる簡単なパス図を作成します.



この状態を式で示すと次のようになります. 潜在変数 Conservative は X で示します.

$$x_1 = \alpha_1 + X\beta_1 + e.x_1$$
$$x_2 = \alpha_2 + X\beta_2 + e.x_2$$
$$x_3 = \alpha_3 + X\beta_3 + e.x_3$$
$$x_4 = \alpha_4 + X\beta_4 + e.x_4$$

ここで我々は次の同時分布を考えます.

$$(X, x_1, x_2, x_3, x_4, e.x_1, e.x_2, e.x_3, e.x_4,)$$

先の図の状態ではこれらの分布は i.i.d であり, その平均ベクトルを μ, 共分散行列を $\mathbf{\Sigma}$ と表現することにします.

ここでモデルを推定すると, 次のような結果を得ます.

```
. sem (Conservative->x1-x4)
```

```
(7186 observations with missing values excluded)
Endogenous variables
Measurement:  x1 x2 x3 x4
Exogenous variables
Latent:       Conservative
Fitting target model:
Structural equation model                       Number of obs     =      1,799
Estimation method  = ml
Log likelihood     = -7571.5183
 ( 1)  [x1]Conservative = 1
```

	OIM					
	Coef.	Std. Err.	z	P>|z|	[95% Conf.	Interval]
Measurement						
x1 <-						
Conservative	1	(constrained)				
_cons	2.328516	.0241199	96.54	0.000	2.281242	2.37579
x2 <-						
Conservative	.8304375	.0561603	14.79	0.000	.7203653	.9405096
_cons	1.61423	.0188197	85.77	0.000	1.577344	1.651116
x3 <-						
Conservative	1.079912	.0653285	16.53	0.000	.9518709	1.207954
_cons	1.416342	.0157728	89.80	0.000	1.385428	1.447257
x4 <-						
Conservative	.9371644	.0569076	16.47	0.000	.8256274	1.048701
_cons	1.3602	.014713	92.45	0.000	1.331363	1.389037
var(e.x1)	.8117739	.0299674			.7551133	.872686
var(e.x2)	.4752263	.0179306			.441351	.5117016
var(e.x3)	.1736938	.0119274			.1518213	.1987173
var(e.x4)	.1831841	.009878			.1648115	.2036047
var(Conservative)	.2348338	.0255071			.1898038	.2905469

```
LR test of model vs. saturated: chi2(2)   =     55.29, Prob > chi2 = 0.0000
```

- 推定結果の各ブロックは先に示した回帰モデルの係数と誤差分散.
- 推定結果の一番下にある尤度比検定はモデルの当てはまりの良さに関する仮説検定.
- 推定したモデルの共分散行列 $\mathbf{\Sigma}$ について考えると, 共分散の存在は仮定していません.
- よって, 次のような制約条件が設定されている状態です.

- 誤差項の共分散

$$\begin{aligned}
\sigma_{e.x_1,e.x_2} &= \sigma_{e.x_2,e.x_1} = 0 \\
\sigma_{e.x_1,e.x_3} &= \sigma_{e.x_3,e.x_1} = 0 \\
\sigma_{e.x_1,e.x_4} &= \sigma_{e.x_4,e.x_1} = 0 \\
\sigma_{e.x_2,e.x_3} &= \sigma_{e.x_3,e.x_2} = 0 \\
\sigma_{e.x_2,e.x_4} &= \sigma_{e.x_4,e.x_2} = 0 \\
\sigma_{e.x_3,e.x_4} &= \sigma_{e.x_4,e.x_3} = 0
\end{aligned} \tag{1.1}$$

- 潜在変数と誤差項の共分散

$$\begin{aligned}
\sigma_{X,e.x_1} &= \sigma_{e.x_1,X} = 0 \\
\sigma_{X,e.x_2} &= \sigma_{e.x_2,X} = 0 \\
\sigma_{X,e.x_3} &= \sigma_{e.x_3,X} = 0 \\
\sigma_{X,e.x_4} &= \sigma_{e.x_4,X} = 0
\end{aligned}$$

- 平均ベクトル μ

$$\begin{aligned}
\mu_X &= 0 \\
\mu_{e.x_1} &= 0 \\
\mu_{e.x_2} &= 0 \\
\mu_{e.x_3} &= 0 \\
\mu_{e.x_4} &= 0
\end{aligned}$$

尤度比検定

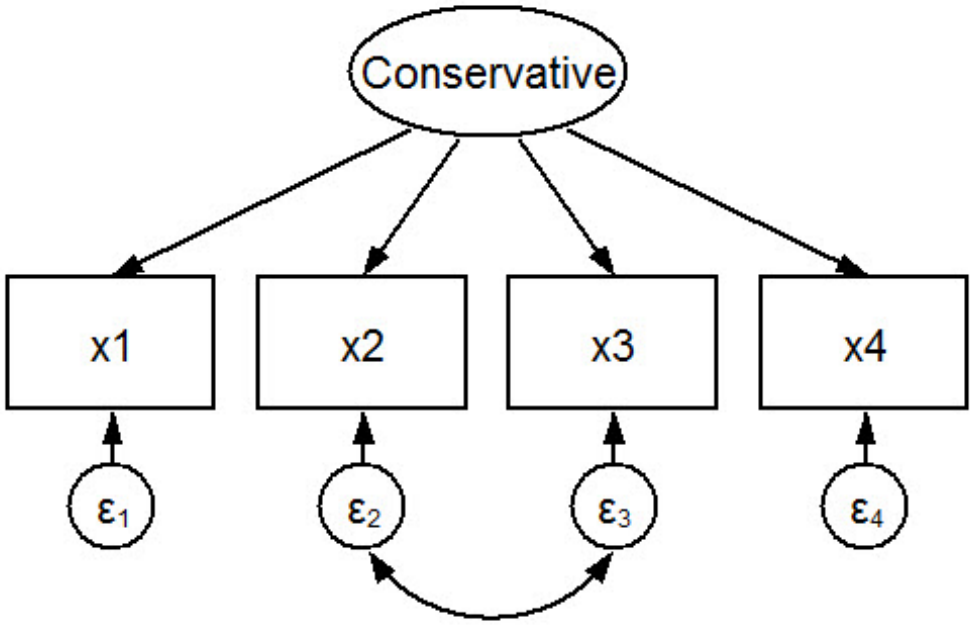
- 共分散の 0 制約を外したモデルが saturated モデルです. つまり, 潜在変数の共分散情報をすべて持っているモデルということになります.
- この他に baseline モデルというものがあります.
- baseline モデルは観測可能な変数の平均と共分散, さらに外生変数が存在する場合はその共分散行列を推定します.
- ただし, 潜在変数との共分散は考えません.

相関の設定

例えば, ここで変数 x_1 と x_2 の間に相関を考えてパス図を次のように更新します.
この時, 先の誤差項における制約の内,

$$\sigma_{e.x_2,e.x_3} = \sigma_{e.x_3,e.x_2} = 0$$

を除外することになります.



参考) 尤度比検定

- Stata は推定結果の表の一番下に尤度比検定の結果を表示しています.
- できるだけ少ない因子で, モデルの分散共分散構造を表現したいという立場から, 帰無仮説は前述のように“構築した SEM モデルは共分散構造を完全に表現できる”です.
- 仮説検定の結果から帰無仮説は棄却され, 改良の余地があることが分かります.
- Stata の PDF マニュアル [SEM] の情報を要約して, ここで行われている尤度比検定の詳細を説明します.

最初に構築したモデルの自由度を df_m,saturated モデルの自由度を df_s と表現することにします. この時,

$$\mathrm{df}_s = \binom{p+q+1}{2} + p + q$$

p は観測可能な内生変数 (質問) の数, q は観測可能な外生変数の数 (ここではゼロ) です.

baseline モデルの自由度は p の数に依存し, 次のようになります.

$$\mathrm{df}_b = \begin{cases} 2q, & \text{if } p = 0 \\ 2p + q + \binom{q+1}{2}, & \text{if } p > 0 \end{cases}$$

モデルの推定手法として最尤法 (ml) や欠損値を考慮した最尤法 (mlmv) を利用する場合の尤度比検定 (baseline モデルと saturated モデル) の検定統計量は次のようになります.

$$\chi^2_{bs} = 2\left(\log L_s - \log L_b\right)$$

自由度は df_{bs} =df_s−df_b. 一方, 構築したモデルと saturated モデルの場合は,

$$\chi^2_{ms} = 2\left\{\log L_s - \log L\left(\hat{\theta}\right)\right\}$$

自由度は df_{ms} =df_s−df_m.

1.2 コマンド入力と SEM ビルダー

Stata では SEM ビルダーでグラフィカルにパス図を作成してモデル推定する方法と, コマンド入力で直接モデル推定を行う方法があります, それぞれ一長一短がありますので, ここでまとめておきます.

1. パス図を利用する場合, 推定結果は係数としてパス図に表示され, 同時に, Stata の Results ウィンドウにも表示されます.
2. コマンドで直接, モデル推定した場合, 推定結果の表だけを表示します.
3. SEM ビルダーに比べ, コマンド入力の方がモデル推定は素早く行えます.
4. コマンド作成の場合, そのコマンドを do ファイルとして保存できます. 従って, エラーが発生した時の原因の調査が簡単です.

コマンド入力の詳細

1. パス図では観測可能な変数は矩形, 潜在変数は円 (楕円) で表現します.

   一般的な Stata のルールとして観測可能な変数は小文字, 潜在変数は先頭または全ての文字を大文字で表記します.

2. コマンドが横に長くなる場合は /// を利用します.

   ```
   . sem (x1<-X) (x2<-X) (x3<-X) (x4<-X)
   ```

   または

   ```
   .sem  (x1<-X)  (x2<-X)    ///
         (x3<-X)  (x4<-X)
   ```

   または,

   ```
   .sem  (x1<-X)    ///
         (x2<-X)    ///
         (x3<-X)    ///
         (x4<-X)
   ```

3. 矢印の方向に制限はありません.

   ```
   (x1<-X)
   (X->x1)
   ```

4. 変数は並列で表記できます.

   ```
   (X->x1 x2 x3 x4)
   ```

次のように記述できます.

```
(X->x1) (X->x2) (X->x3) (X->x4)
```

または,

```
(x1<-X) (x2<-X) (x3<-X) (x4<-X)
```

```
(x1 x2 x3 x4 <-X)
```

少し複雑なモデルの場合として,

```
(X Y ->x1 x2 x3) (X->x4 x5) (Y->x6 x7)
```

これは次のようも記述できます.

```
(X->x1 x2 x3 x4 x5)     ///
(Y->x1 x2 x3 x6 x7)
```

おなじく,

```
(X->x1) (X->x2) (X->x3) (X->x4) (X->x5)     ///
(Y->x1) (Y->x2) (Y->x3) (Y->x6) (Y->x7)
```

5. パス図では誤差項を配置する必要がありますが, コマンド入力では省略できます.

```
(x1<-X) (x2<-X) (x3<-X) (x4<-X)
```

これを敢えて明示的に書けば,

```
(x1<-X e.x1)     ///
(x2<-X e.x2)     ///
(x3<-X e.x3)     ///
(x4<-X e.x4)
```

パス係数に 1 という制約を掛ける場合は,

```
(x1<-X e.x1@1)     ///
(x2<-X e.x2@1)     ///
(x3<-X e.x3@1)     ///
(x4<-X e.x4@1)
```

もちろん, これは次のようにも書けます.

```
(x1<-X@1)     ///
(x2<-X@1)     ///
(x3<-X@1)     ///
(x4<-X@1)
```

6. 制約の無い場合,

```
(x1<-X) (x2<-X) (x3<-X) (x4<-X)
```

ここで x2<-X のパス係数を 2 とする場合は,

```
(x1<-X) (x2<-X@2) (x3<-X) (x4<-X)
```

x2<-X と x3<-X の係数は等しいという制約を掛ける場合は,

```
(x1<-X) (x2<-X@b) (x3<-X@b) (x4<-X)
```

7. 共分散の存在を仮定する場合の曲線矢印を意図する場合は,

```
(x1 x2 x3 x4 <-X), cov(e.x2*e.x3)
```

e.x2*e.x3 と e.x3*e.x4 の 2 つのセットで考える場合は,

```
(x1 x2 x3 x4 <-X), cov(e.x2*e.x3 e.x3*e.x4)
```

または,

```
(x1 x2 x3 x4 <-X), cov(e.x2*e.x3) cov(e.x3*e.x4)
```

1.3 適合度の補足

- 次のコマンドでモデルを推定し, 適合度検定の情報を補足します.
- あくまで, 検定の意味を確認するための操作です. モデルの本質的な意味を考慮しての操作ではありません.

```
. sem (Conservative->x1-x4), cov(e.x2*e.x3)
```

(推定結果は省略します)

```
. estat gof,stats(all)
```

Fit statistic	Value	Description
Likelihood ratio		
chi2_ms(1)	38.830	model vs. saturated
p > chi2	0.000	
chi2_bs(6)	1461.871	baseline vs. saturated
p > chi2	0.000	
Population error		
RMSEA	0.145	Root mean squared error of approximation
90% CI, lower bound	0.108	
upper bound	0.186	
pclose	0.000	Probability RMSEA <= 0.05
Information criteria		
AIC	15152.573	Akaike´s information criterion
BIC	15224.008	Bayesian information criterion
Baseline comparison		
CFI	0.974	Comparative fit index
TLI	0.844	Tucker-Lewis index
Size of residuals		
SRMR	0.032	Standardized root mean squared residual
CD	0.835	Coefficient of determination

- RMSEA の項目について補足します.
- 先の例で, 一般的に 0.05 で良く, 0.08 でほどほどに良いフィットであることを述べました.
- したがって, RMSEA=0.145 という結果は好ましいものではありません. その下にある 90% 信頼区間については次のように考えます.

- lower bound:これが 0.05 よりも小ならば, 「フィットが良い」を棄却できない. 従ってこの例では棄却されます.
- upper bound:0.1 よりも大ならば, 「フィットが悪い」が棄却できない. ここでは棄却できません.
- pclose は RMSEA が 0.05 以下になる確率. ここではゼロに近いので, フィットが悪いことと整合的です.

1.4　2 ファクターモデルのための準備

- ここまでは Stata の PDF マニュアル [SEM] の内容を利用して第一回の解説ではカバーできなかった基本的な事柄について解説しました.
- ここからは話を STATA Press の *Discovering Structural Equation Modeling Using Stata, Revised Edition* に戻して, モデルの構築を進めます.
- サンプルデータである nlsy97cfa.dta には Conservative(保守性) に関する質問項目の他に, 鬱 (Depression) の心理状態に関する調査項目が含まれています.
- 最終的には観測できない保守性という因子と, 同じく観測できない鬱という因子の関係をモデル化します.
- 鬱症状の重い人ほど, 保守性が強くなる, というようなことが手元のデータから言えるのでしょうか?

そのための第一歩として, 鬱に関連する質問項目ついて概観することから作業を始めます.

```
. codebook x11-x13,compact
```

```
Variable    Obs Unique      Mean  Min  Max  Label
──────────────────────────────────────────────────────────────────────────────
x11        7295      4  3.223715    1    4  HOW OFT R FELT DOWN OR BLUE 2008
x12        7397      4  2.232932    1    4  HOW OFT R BEEN HAPPY PERSON 2008
x13        7291      4  3.655328    1    4  HOW OFT R DEPRESSED LAST MONTH 2008
──────────────────────────────────────────────────────────────────────────────
```

- 被験者の数は 7,291 から 7,397 人で, 保守性に関する調査項目よりもかなり多いことが分かります.

鬱症状に関するモデリング

- いきなり 2 ファクターモデルを推定するような事は避けてください.
- まずは, 個別に SEM による単独のモデリングを行います.
- 今, 3 つの項目 (質問) がありますので, その分散共分散行列の要素は $3\,(3+1)\,/2 = 6$ 個です.

- 具体的なモデルは,

$$x_{11} = \alpha_{11} + D\beta_{11} + e.x_{11}$$
$$x_{12} = \alpha_{12} + D\beta_{12} + e.x_{12}$$
$$x_{13} = \alpha_{13} + D\beta_{13} + e.x_{13}$$

- 3 つの質問からなるこのモデルのパラメータは 6 個で, 丁度識別の状態になっています.
- 丁度識別の場合, 尤度比検定の自由度は 0 になり, 構築したモデルが saturated モデルとまったく同じ情報を持っているので, 仮説検定は実行できません.
- 質問が 2 つしかないと, 推定自体が実行できません.

3 つの質問項目からなるモデルを推定します. 丁度識別の場合, 推定後に行う `estat gof` や `estat mindices` などの検定を行っても意味はありません.

```
. sem (Depress -> x11-x13)
```

(推定結果は省略します)

```
. sem,standardized
```

```
Structural equation model                       Number of obs     =      7,183
Estimation method  = ml
Log likelihood     = -18464.972
 ( 1)  [x11]Depress = 1
```

Standardized	Coef.	OIM Std. Err.	z	P>|z|	[95% Conf.	Interval]
Measurement						
x11 <-						
Depress	.8130901	.0101864	79.82	0.000	.7931251	.8330551
_cons	4.851163	.0421589	115.07	0.000	4.768533	4.933793
x12 <-						
Depress	-.6088417	.0102152	-59.60	0.000	-.6288631	-.5888203
_cons	3.435663	.0309978	110.84	0.000	3.374909	3.496418
x13 <-						
Depress	.654818	.010117	64.72	0.000	.6349891	.6746469
_cons	6.159645	.0527282	116.82	0.000	6.0563	6.26299
var(e.x11)	.3388845	.0165649			.3079245	.3729572
var(e.x12)	.6293117	.0124389			.6053982	.6541699
var(e.x13)	.5712134	.0132496			.5458261	.5977814
var(Depress)	1	.			.	.

```
LR test of model vs. saturated: chi2(0)   =      0.00, Prob > chi2 =      .
```

- x12 は変数の値が大きいほど, 鬱とは反対の状態を指しますので, 符号が逆転しています.
- 推定結果の解釈を簡単に行いたい, という意図がある時はリバースコーディングします. モデル推定上は特に問題にはなりません.
- 潜在変数 Depress の信頼性は $\rho = 0.74$ です. 信頼性を計算する場合は, 次のコマンドで潜在変数の分散を 1 に固定してモデルを再推定します.

```
. sem (Depress -> x11-x13),var(Depress@1)
```

- 信頼性の計算においては推定値の絶対値を利用します.

1.5　推定手法と欠損値

- 最終的に推定する 2 ファクターモデルでは, 保守性と鬱の質問のうち, 片方の質問にすべて答えていないような標本は除いてモデル推定を行いたいと考えています.
- そこで, 推定手法の特徴と, 具体的なコマンドについてここで考察しておきます.

Option1. sem (Depress -> x11-x13)

- これはデフォルトのコマンドです. この場合, Depress の項目について欠損値があるとリストワイズな削除を行います.
- その結果, モデル推定には 7,183 人分のデータ (3 問に回答) を利用します.
- Conservative への回答状況は考慮しません.

Option 2. sem (Depress -> x11-x13),method(mlmv) allmiss

- 少なくとも 1 つの質問には答えている 7,429 人分のデータを利用します.
- Conservative への回答状況は考慮しません.
- `allmiss` オプションは全ての種類の欠損値に対応します.

Option 3. sem (Depress -> x11-x13) if !missing(x1, x2, x3,x4,x5, x6, x7, x8, x9)

- Depress の項目について欠損値があるとリストワイズな削除を行います.
- Conservative についても同様にリストワイズ削除を行います.
- 標本数は 1,460 人です.

Option 4. sem (Depress -> x11-x13) if x1 !=. | x2 !=. |x3 !=. |x4 !=. |x5 !=. |

x6 !=. |x7 !=. |x8 !=. |x9 !=.,method(mlmv) allmiss

- Depress のうちの少なくとも一つ, 同じく, Conservative の少なくとも一つに回答した 1,752 人を利用.

以上