LightStone

Stata で簡単に試せる SEM

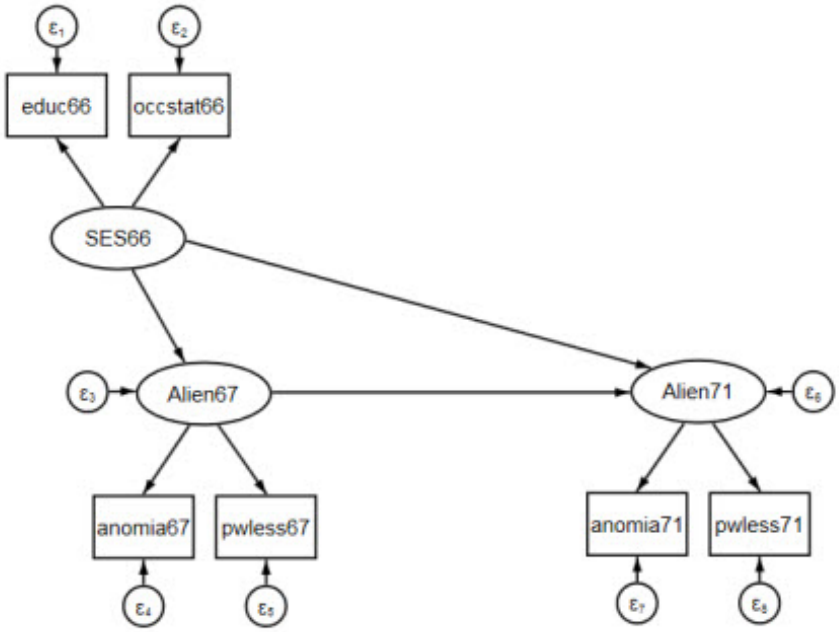
第4回　構造方程式モデリング

SEM の第 4 回目です．今回の目的はこれまでの知識を利用して構造方程式モデルを推定することです．Alan C. Acock, 2013.　*Discovering Structural Equation Modeling Using Stata, Revised Edition*, Stata Press の第 3 章を利用して説明します．

第4章　構造方程式モデリング

最初に今回の目的であるモデルの特徴を, 2 ファクタモデルと比較することで明らかにする.[1]

- 2 ファクタモデルでは 2 つの潜在変数 Depress と Conservative の間に相関を仮定した.
- 今回は 3 ファクタを考える. しかし, この 3 つのファクタ間には相関ではなく, 時間軸で考えたときの一方から他方への影響を考える.



この図から潜在変数 SES66 が Alien67 と Alien71 に直接影響を与え, 一部は Alien67 を経由して Alien71 に, 間接的に影響を与えていることが分かる.

データの構成

Wheaton et al. (1977) 以前は社会経済的地位と, 疎外感と言った心理的な事象を, 観測可能な変数でモデリングすることは困難であるとされていた. Wheaton らはその関係を構造方程式モデリングを利用して, 統計的に有意な結果を得る事に成功した.

- Wheaton et al. (1977) は 1967-1971 年の 5 年間のデータを利用して疎外感をモデル化した.

[1] Alan C. Acock, 2013. *Discovering Structural Equation Modeling Using Stata, Revised Edition*, Stata Press の第 3 章を利用して説明します.

- SES66 は社会経済的地位を示す変数で, ここでは潜在変数とする.
- educ66 と occstat66 は計測した変数で学歴と職業的地位である.
- Alien67 と Alien71 はそれぞれの調査時点における疎外感で潜在変数とする.
- 変数 anonima は健忘性失語症, pwless(powerlessness) は無力感.
- ここでは潜在変数間の矢印の向きが重要である.

4.1　SES と疎外感の関係

今回利用するデータは sem_sm2.dta.

```
.use sem_sm2,clear
```

- このデータはアンケートの元データではない. 集計済みのデータ (平均, 標準偏差, 分散共分散) である.
- 前回の例題までは集計前の生データを利用して分析を行った.
- この時のデータ形式を Stata では Summary Statistics Data(SSD) と呼ぶが, その詳細については最後に紹介する.
- 最初に示したパス図を作成し,「 標準化係数」を利用する方法でモデル推定する.

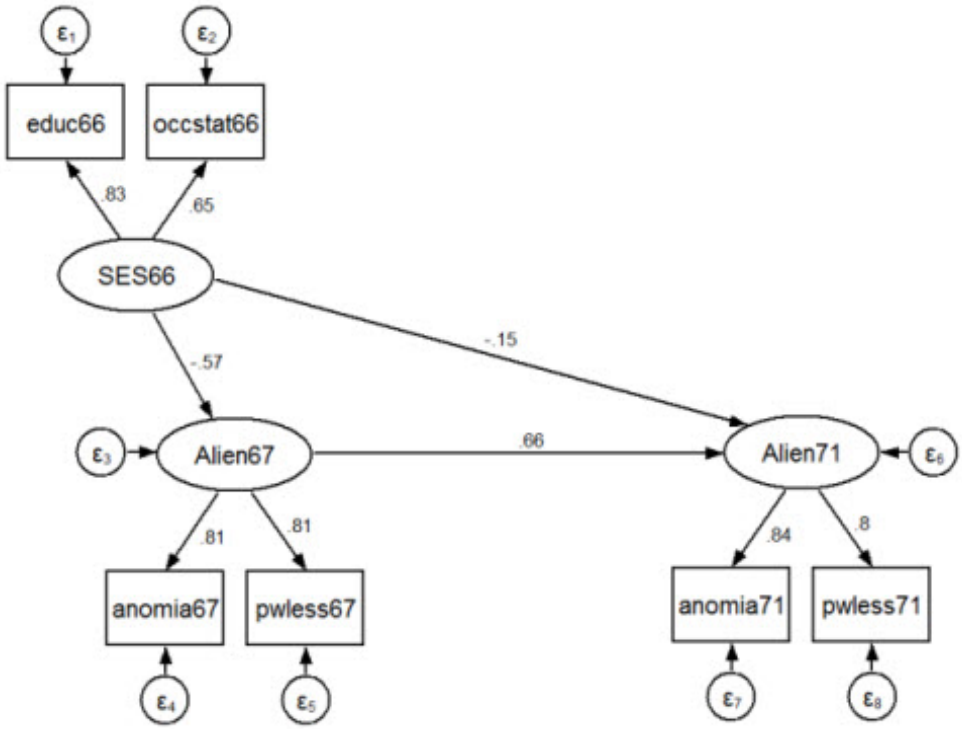
```
Endogenous variables
Measurement:  educ66 occstat66 anomia67 pwless67 anomia71 pwless71
Latent:       Alien67 Alien71
Exogenous variables
Latent:       SES66
Fitting target model:
(省略)
Structural equation model                       Number of obs     =        932
Estimation method  = ml
Log likelihood     = -15246.469
 ( 1)  [anomia67]Alien67 = 1
 ( 2)  [anomia71]Alien71 = 1
 ( 3)  [educ66]SES66 = 1
```

Standardized	Coef.	OIM Std. Err.	z	P>|z|	[95% Conf.	Interval]
Structural						
Alien67 <-						
SES66	-.5668218	.0344036	-16.48	0.000	-.6342517	-.4993919
Alien71 <-						
Alien67	.6630088	.0396724	16.71	0.000	.5852523	.7407654
SES66	-.151492	.0458162	-3.31	0.001	-.24129	-.061694
Measurement						
educ66 <-						
SES66	.8326718	.031738	26.24	0.000	.7704664	.8948772
_cons	3.518017	.0878219	40.06	0.000	3.345889	3.690145
occstat66 <-						
SES66	.6485148	.0301669	21.50	0.000	.5893887	.707641
_cons	1.767678	.0524337	33.71	0.000	1.66491	1.870446
anomia67 <-						
Alien67	.812882	.0194328	41.83	0.000	.7747943	.8509697
_cons	3.95852	.097363	40.66	0.000	3.767692	4.149347
pwless67 <-						
Alien67	.811926	.0194466	41.75	0.000	.7738113	.8500406
_cons	4.796692	.1158294	41.41	0.000	4.56967	5.023713
anomia71 <-						
Alien71	.8395125	.0193263	43.44	0.000	.8016337	.8773913
_cons	3.993669	.09813	40.70	0.000	3.801338	4.186
pwless71 <-						
Alien71	.798082	.0198613	40.18	0.000	.7591546	.8370095
_cons	4.717723	.1140761	41.36	0.000	4.494137	4.941308
var(e.educ66)	.3066577	.0528548			.2187474	.4298974
var(e.occstat66)	.5794285	.0391274			.5075984	.6614233
var(e.anomia67)	.3392229	.0315932			.2826241	.4071562
var(e.pwless67)	.3407762	.0315784			.2841788	.4086457
var(e.anomia71)	.2952187	.0324493			.2380034	.3661885
var(e.pwless71)	.3630651	.0317019			.3059565	.4308333
var(e.Alien67)	.6787131	.0390015			.6064191	.7596255
var(e.Alien71)	.4236057	.0345717			.360988	.4970851
var(SES66)	1	.			.	.

```
LR test of model vs. saturated: chi2(6)   =     71.62, Prob > chi2 = 0.0000
```

- 回帰係数だけを表示した時のパス図は次のようになる.

- 誤差項の分散と内生変数の平均は表示しない.



推定結果を順番に見てゆくと, 次のようなことが分かる.

- 標準化係数を利用したので, 3 つの潜在変数のブロックでそれぞれ一つの観測可能な変数の分散が 1 になっている.

```
( 1) [anomia67]Alien67 = 1
( 2) [anomia71]Alien71 = 1
( 3) [educ66]SES66 = 1
```

- 潜在変数が観測可能な変数に与える影響の強さを係数で見ると 0.65 以上, 最大で 0.85 となっており, 潜在変数から観測可能な変数への影響の強いことが分かる.
- 推定結果の表の Structural 部分 (潜在変数の関係) を見ると, 一番影響が強いのは Alien67→Alien71 で, $\beta = 0.66$ で有意である.

各方程式の適合度

パス図における回帰係数を示す式の適合を個別度に考察する.

```
. estat eqgof
```

```
 Equation-level goodness of fit
```

depvars	fitted	Variance predicted	residual	R-squared	mc	mc2
observed						
educ66	9.599689	6.65587	2.943819	.6933423	.8326718	.6933423
occstat66	449.8053	189.1753	260.63	.4205715	.6485148	.4205715
anomia67	11.8209	7.810982	4.009921	.6607771	.812882	.6607771
pwless67	9.353552	6.166084	3.187468	.6592238	.811926	.6592238
anomia71	12.51815	8.822558	3.695593	.7047813	.8395125	.7047813
pwless71	9.974882	6.35335	3.621531	.6369349	.798082	.6369349
latent						
Alien67	7.810982	2.509567	5.301416	.3212869	.5668218	.3212869
Alien71	8.822558	5.085272	3.737286	.5763943	.7592064	.5763943
overall				.7784845		

```
mc  = correlation between depvar and its prediction
mc2 = mc^2 is the Bentler-Raykov squared multiple correlation coefficient
```

- 潜在変数 Alien67 の変動の 32.1%, 同じく潜在変数 Alien71 の 57.6% をモデルで説明できている.
- SEM を利用する以前の研究では, この関係は統計的に安定的なものではないとされていた.
- Alien67→Alien71 の回帰係数 $\beta = 0.66$ は有意であり, 決して小さい値ではない.

SEM 以前と比べ, anonima(健忘性失語症) と無力感を疎外感 (Alien) に関する変数として位置づけ, これら以外の要因を誤差項として設定したことが大きな違いをもたらしたと考えられている.

モデル全体での評価

次はモデル全体での適合度を評価する.

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

Fit statistic	Value	Description
Likelihood ratio		
chi2_ms(6)	71.621	model vs. saturated
p > chi2	0.000	
chi2_bs(15)	2134.080	baseline vs. saturated
p > chi2	0.000	
Population error		
RMSEA	0.108	Root mean squared error of approximation
90% CI, lower bound	0.087	
upper bound	0.131	
pclose	0.000	Probability RMSEA <= 0.05
Information criteria		
AIC	30534.938	Akaike´s information criterion
BIC	30636.522	Bayesian information criterion
Baseline comparison		
CFI	0.969	Comparative fit index
TLI	0.923	Tucker-Lewis index
Size of residuals		
SRMR	0.021	Standardized root mean squared residual
CD	0.778	Coefficient of determination

- カイ二乗検定 chi2_ms(6) の帰無仮説は, 「推定したモデルは変数の分散共分散の情報を完全に表現している」である. ここでは帰無仮説が棄却されているので, モデルに改良の余地があることが分かる.
- RMSE は 0.11 で目安の 0.05 を越えており, フィットは基準ほどは良くない.
- CFI は目安の 0.95 を越えた 0.97 で, 基準よりも良いことが分かる.

モデルの改善可能性

MI によるモデルを改善可能性に関する検定を実行する.

```
. estat mindices
```

Modification indices

	MI	df	P>MI	EPC	Standard EPC
Measurement					
educ66 <-					
anomia67	4.415	1	0.04	.1055965	.1171781
pwless67	6.816	1	0.01	-.1469371	-.1450411
anomia67 <-					
educ66	5.627	1	0.02	.0935048	.0842631
anomia71	51.977	1	0.00	.3906425	.4019984
pwless71	32.517	1	0.00	-.2969297	-.2727609
pwless67 <-					
educ66	6.441	1	0.01	-.0889042	-.0900664
anomia71	41.618	1	0.00	-.3106995	-.3594367
pwless71	23.622	1	0.00	.2249714	.2323233
anomia71 <-					
anomia67	58.768	1	0.00	.429437	.4173061
pwless67	38.142	1	0.00	-.3873066	-.3347904
pwless71 <-					
anomia67	46.188	1	0.00	-.3308484	-.3601641
pwless67	27.760	1	0.00	.2871709	.2780833
cov(e.educ66,e.anomia67)	6.063	1	0.01	.5527612	.1608845
cov(e.educ66,e.pwless67)	7.752	1	0.01	-.5557802	-.1814365
cov(e.anomia67,e.anomia71)	63.786	1	0.00	1.951578	.5069627
cov(e.anomia67,e.pwless71)	49.892	1	0.00	-1.506704	-.3953794
cov(e.pwless67,e.anomia71)	49.876	1	0.00	-1.534199	-.4470094
cov(e.pwless67,e.pwless71)	37.357	1	0.00	1.159123	.341162

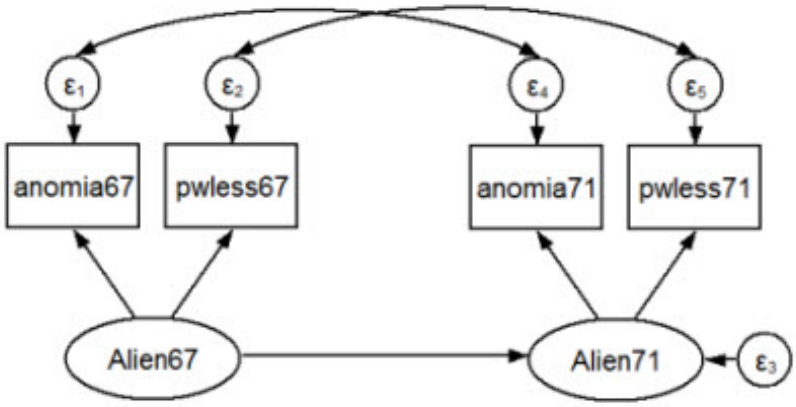
EPC = expected parameter change

- MI の大きさはカイ二乗値の差を示している
- ただし, 分散に関するパラメータとして不合理的なものを利用しないこと
- 例えば, anomia71 が anomia67 に影響を与えることはないので, この両者の間に相関を設定してはいけない.

モデルの編集

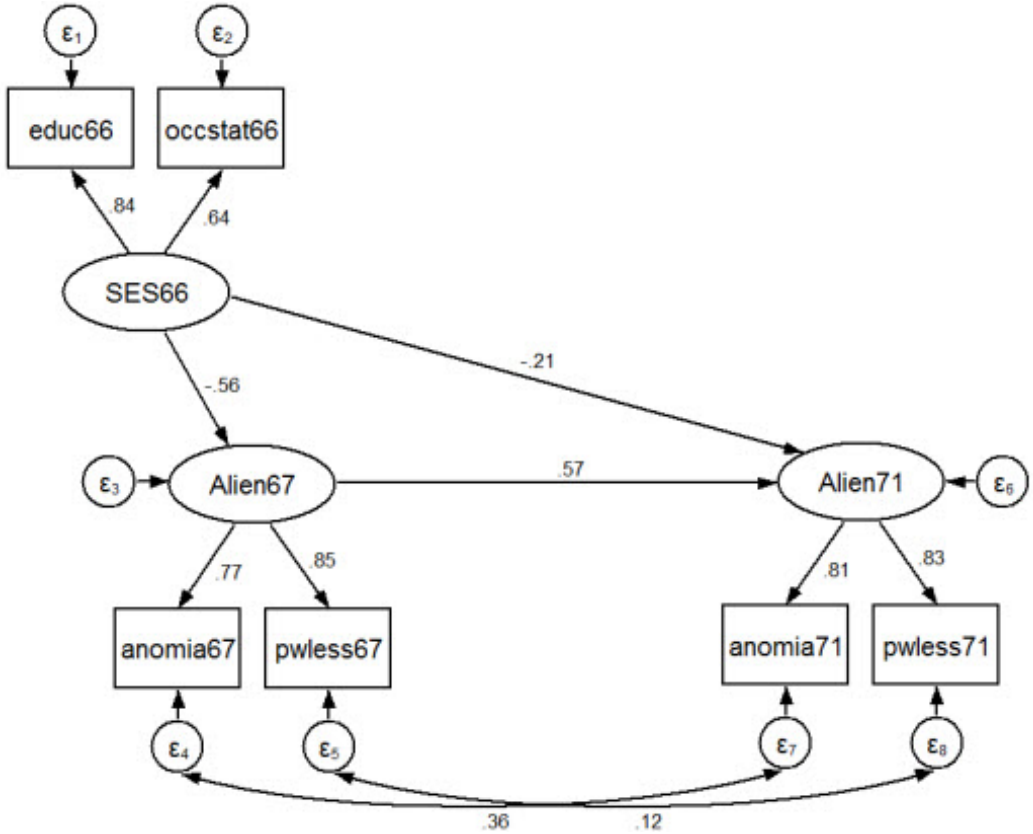
- 誤差項に着目する
- 観測可能な anomia や powerless 以外の, 観測できない, または, 入手できない共通要因が存在し, それは誤差項に含まれているものとする.
- その要因が anomia と powerless の症状に, 各年で影響を与えていると考える.

つまり,



- 各要素に共通の要因を考えると, その結果として回帰係数の影響は小さくなることが予想される

間接効果



- SES66 が直接 Alien71 に与える効果は「直接効果」と呼ぶ
- SES66 から Alien67 を介して, Alien71 に与える効果を「間接効果」と呼ぶ
- パス図を見ると Alien67 から Alien71 の係数は小さくなっている (0.66 → 0.57)

- `estat gof,stats(all)` でモデル全体の適合度を比較すると, フィットは改善されていることが分かる

改善前	改善後
$\chi^2(6) = 71.62, p = 0.000$	$\chi^2(4) = 4.77, p = 0.31$
RMSE=0.108	RMSE=0.014
CFI=0.97	CFI=1.000
MI を複数表示	十分改善されているので非表示

- 次のコマンドで直接効果と間接効果を求める.

```
. estat teffects,nodirect
```

(コマンド実行結果の一部)

```
Structural   |

 Alien67
      SES66 |          0  (no path)
------------+----------------------------------------------------------------
 Alien71
    Alien67 |          0  (no path)
      SES66 | -.3491338   .0412546    -8.46   0.000    -.4299914   -.2682762
```

- 出力された Indirect effects の表を見ると, SES66 から Alien67 への間接効果が-3.49 で有意であることが分かる.

簡単なまとめ

- 潜在変数間における影響の与えた方を考慮したモデリングを行った
- 間接効果と直接効果による潜在変数の考察を行った

4.2 集計したデータによる分析

SEM によるデータ分析を行う場合, 次のような理由で既に集計されたデータを利用するような事がある.

- 個人の回答を公表したくない
- 調査に参加した個人が特定されるような可能性は残したくない

このような場合に備えて, Stata では集約した情報から SEM を行う機能を備えている.

- 集計済みデータのことを Stata では SSD(Summary statistics data) と呼ぶ.
- ただし, SSD は sem コマンドの場合だけに対応しており, gsem には対応してない.

例題

1. 3 つの変数 x1,x2,x3 があるものとする. アンケートに 3 つの質問があり, それぞれに選択肢が 5 個あり, その選択肢の数字には大小関係があるような状況とする.
2. アンケートに回答してくれた人は 74 人とする.
3. 回答の分散共分散行列は次のようになった.

$$\begin{matrix} 33.4722 & & \\ -3.6294 & 0.6043 & \\ 1.0374 & -0.2120 & 0.2118 \end{matrix}$$

4. 各回答の平均は 21.2973, 3.0195, 0.2973 であるとする.

集計データの入力

次に示す手順でデータを入力する.

1. メモリー中のデータをクリアする.

```
.clear all
```

2. 変数名を設定する.

```
.ssd init x1 x2 x3
```

3 番以降のステップに明確な順番はない.

3. データの個数を設定する.

```
.ssd set obs 74
```

4. 分散共分散の情報を入力する.

```
.ssd set cov 33.4722 \ -3.6294 .6043 \ 1.0374 -.2120 .2118
```

相関行列の場合は ssd set cor とする.

5. 平均を入力する.

```
.ssd set means 21.2973 3.0195 .2973
```

6. 設定した変数の情報を任意の時点で確認する場合は次のように入力する.

```
.ssd status
```

7. 入力した情報を具体的に表示する場合は次のように入力する.

```
.ssd list
```

8. タイプミスがあった時は, オプション `replace` を付けて再入力する.

```
.ssd set means 21.2973 3.0195 .2973,replace
```

9. データ入力後は, 通常の場合と同じ手順でデータを保存する.
10. 集計データで SEM を行う場合の手順は, 元データの場合とまったく同じ.

SSD 利用上の注意点

一般的に標準誤差の計算に関して, 計算上の制約が多くなる. SSD で SEM コマンドを実行する時に利用できない機能は次の通り.

1. sem コマンドと `vce(sbentler)` オプションで実行可能な Satorra-Bentler 標準誤差の計算と, Satorra-Bentler スケール化 χ^2 検定は利用できない.
2. sem コマンドと `vce(robust)` オプションで求める堅牢な標準誤差は計算できない.
3. sem コマンドと `vce(cluster` *clustvar*`)` オプションで求めるクラスタ対応の標準誤差は計算できない.
4. svy:プリフィックスと sem コマンドで計算可能なサーベイデータ用の標準誤差は計算できない.
5. sem コマンドと `vce(bootstrap)` または `vce(jackknife)` オプションで計算可能なブートストラップおよびジャックナイフ標準誤差は計算できない.
6. sem コマンドと `vce(opg)` オプションで計算可能な分散共分散推定値は計算できない.
7. 例えば, `[fw=varname]` で利用するような加重は利用できない.
8. if や in を利用して推定に利用する標本を制限することはできない.
9. 欠損値に対応した `method(mlmv)` や漸近分布を利用する `method(adf)` のオプションは利用できない.

実際に今回用いたサンプルデータの情報を確認してみよう.

```
.use sem_sm2,clear
.ssd list
```

データにエディタに入力されていたデータは平均, 標準偏差, 相関行列であることが分かる.

4.3 GSEM

ここまで利用したモデルは線形モデルである. [2]

- ここからは非線形モデルの場合に利用する GSEM コマンドについて説明する.
- GSEM ではプロビット, ロジスティック, 順序プロビット, 多項ロジスティックなどの回帰モデルが利用できる.

最初にデータをダウンロードします.

```
. webuse gsem_1fmm,clear
. sum
```

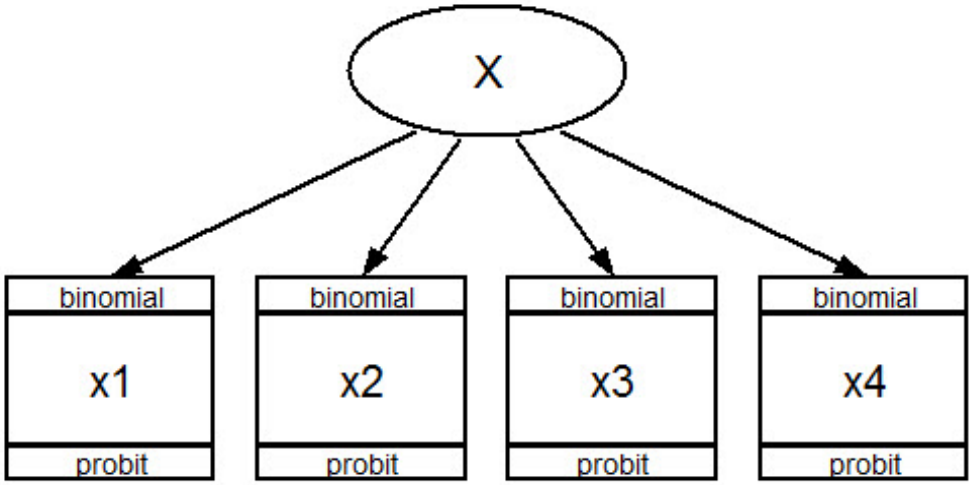
Variable	Obs	Mean	Std. Dev.	Min	Max
x1	123	.4065041	.4931897	0	1
x2	123	.4065041	.4931897	0	1
x3	123	.4227642	.4960191	0	1
x4	123	.3495935	.4787919	0	1
s4	123	690.9837	77.50737	481	885

- このサンプルデータには x1 から x4 までの 4 回の試験に合格した/しないの結果が 0 と 1 で用意されている.
- x1 から x3 までの試験は満点が 100 点未満
- x4 は 725 点未満
- s4 には 4 回目の試験の実際の点数が入っている

プロビットモデルを利用する

x1 から x4 までの試験結果を利用して, 観測できない能力 X をモデル化する. 目的とするパス図は次の通り.

[2] ここからは Stata の英文マニュアルで紹介されている内容を用いて解説します.



- SEM ビルダーで非線形モデルを利用する場合はアイコン 

を最初にクリックする.

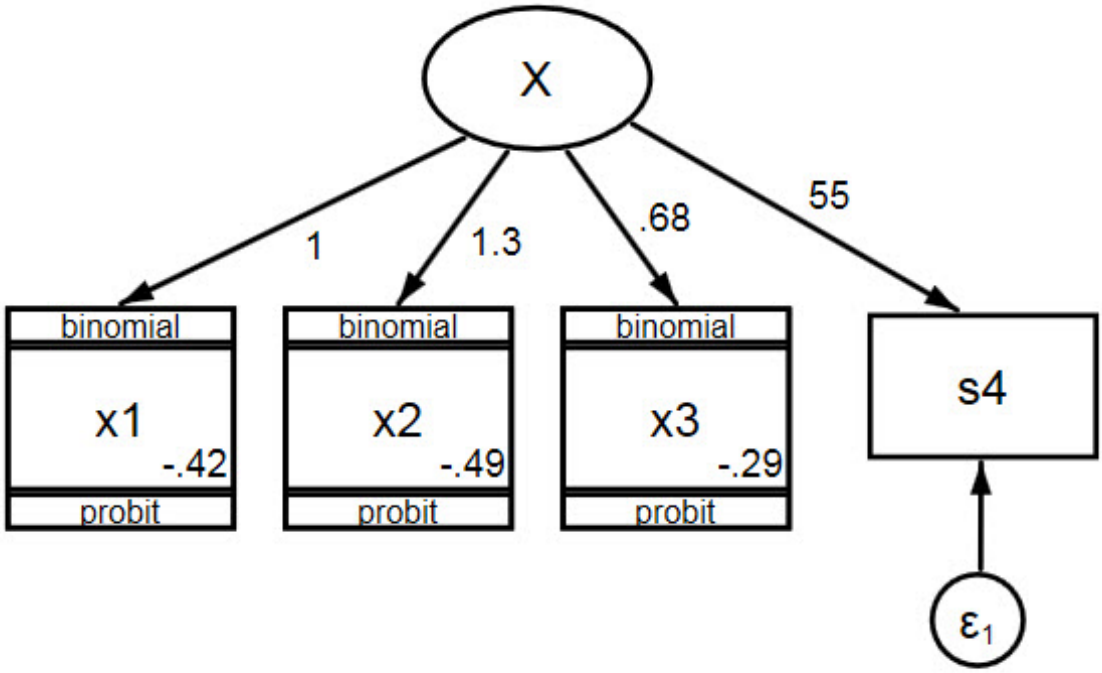
- GSEM の場合には明らかにパス図のデザインが SEM の場合とは異なる.

- しかし, モデルの推定手順自体は同じ.
- 実際に推定した結果を次す. ただし, GSEM では標準化係数という選択肢はない.

```
Fitting fixed-effects model:
Iteration 0:   log likelihood = -329.82091
Iteration 1:   log likelihood = -329.57665
Iteration 2:   log likelihood = -329.57664
(省略します)
Generalized structural equation model           Number of obs     =        123
Response       : x1
Family         : Bernoulli
Link           : probit
Response       : x2
Family         : Bernoulli
Link           : probit
Response       : x3
Family         : Bernoulli
Link           : probit
Response       : x4
Family         : Bernoulli
Link           : probit
Log likelihood = -261.30263
 ( 1)  [x1]X = 1
```

	Coef.	Std. Err.	z	P>|z|	[95% Conf.	Interval]
x1 <-						
X	1	(constrained)				
_cons	-.3666763	.1896773	-1.93	0.053	-.738437	.0050844
x2 <-						
X	1.33293	.4686743	2.84	0.004	.4143455	2.251515
_cons	-.4470271	.2372344	-1.88	0.060	-.911998	.0179438
x3 <-						
X	.6040478	.1908343	3.17	0.002	.2300195	.9780761
_cons	-.2276709	.1439342	-1.58	0.114	-.5097767	.0544349
x4 <-						
X	9.453342	5.151819	1.83	0.067	-.6440375	19.55072
_cons	-4.801027	2.518038	-1.91	0.057	-9.736291	.1342372
var(X)	2.173451	1.044885			.847101	5.576536

- この推定結果から質問項目 x1 から x4 でプロビットモデルが利用されていることが分かる.
- 試験の x4 については手元に実際の得点 s4 がある. したがって, パス図を次のように作り変えてみよう.



このように編集した後で推定した結果を次に示す.

```
Fitting fixed-effects model:
(省略)
Generalized structural equation model           Number of obs     =        123
Response       : x1
Family         : Bernoulli
Link           : probit
Response       : x2
Family         : Bernoulli
Link           : probit
Response       : x3
Family         : Bernoulli
Link           : probit
Response       : s4
Family         : Gaussian
Link           : identity
Log likelihood =  -869.6892
 ( 1)  [x1]X = 1
```

	Coef.	Std. Err.	z	P>|z|	[95% Conf.	Interval]
x1 <-						
X	1	(constrained)				
_cons	-.4171085	.1964736	-2.12	0.034	-.8021896	-.0320274
x2 <-						
X	1.298311	.3280144	3.96	0.000	.6554142	1.941207
_cons	-.4926357	.2387179	-2.06	0.039	-.9605142	-.0247573
x3 <-						
X	.682969	.1747328	3.91	0.000	.3404989	1.025439
_cons	-.2942021	.1575014	-1.87	0.062	-.6028992	.0144949
s4 <-						
X	55.24829	12.19904	4.53	0.000	31.3386	79.15798
_cons	690.9837	6.960106	99.28	0.000	677.3422	704.6253
var(X)	1.854506	.7804393			.812856	4.230998
var(e.s4)	297.8565	408.64			20.24012	4383.299

このように SEM には非線形モデルに対応した GSEM という推定機能があり, 非線形性が想定される様々な調査, 質問パターンに対応している.

Stata 15 の [SEM] マニュアルにある example 30g を利用して解説を行います. ただし, ここではマルチレベル分析の手法に関する説明は行いませんのでご了承ください.