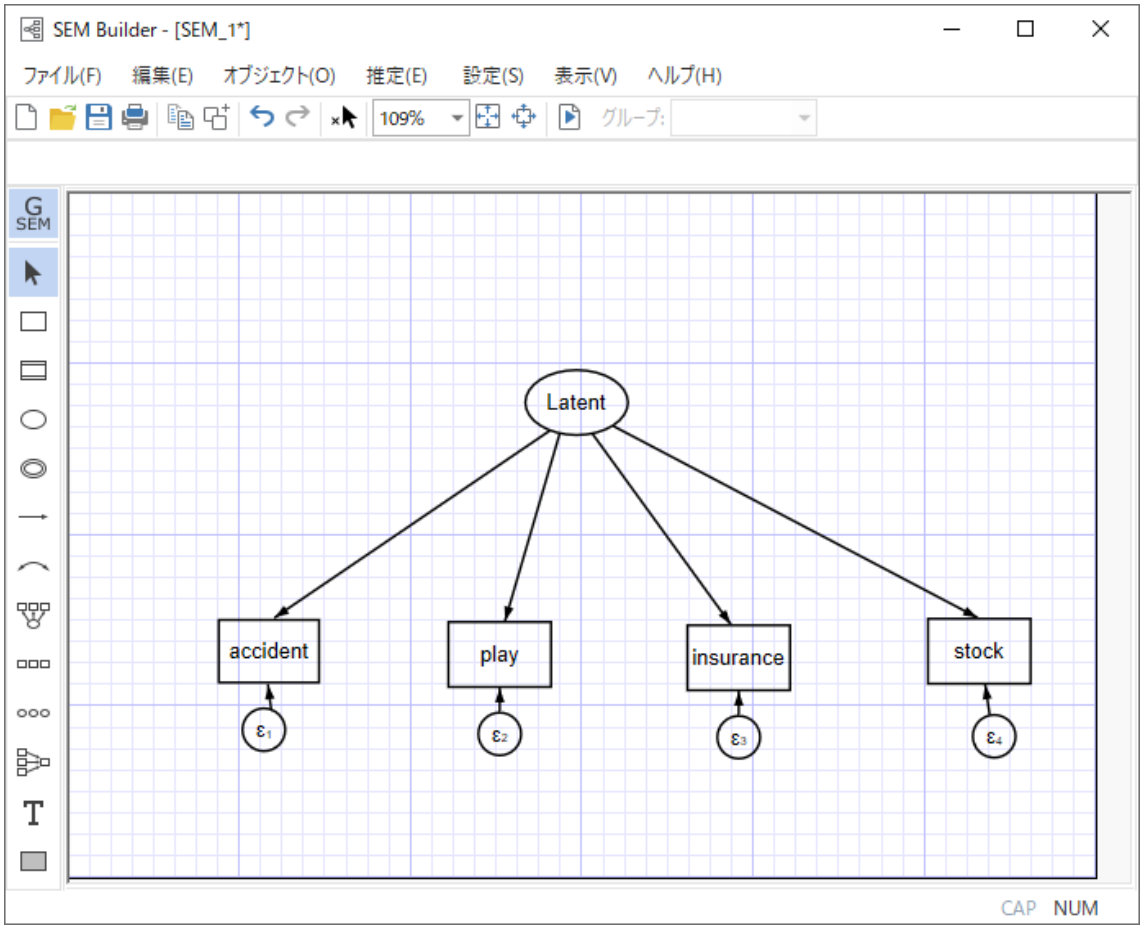
株式会社ライトストーン
分析例題集

潜在クラス分析

1. イントロダクション

潜在クラス分析（LCA）では，母集団にはグループがあり，これらのグループの個人は異なる行動をとると考えます．しかし，このグループを識別する変数はありません．グループは，さまざまな購買嗜好を持つ消費者，さまざまな行動パターンを持つ少年，または健康状態の分類である可能性があります．LCA は，これらの観測できないグループ分けを特定し，標本ごとのグループへの所属や，そのグループの特徴が他のグループとどのように異なるかを知ることができます．

潜在クラスモデルは，(連続ではなく)カテゴリの潜在変数を持つことによって特徴付けられます．カテゴリ潜在変数のレベルは母集団のグループを表し，クラスと呼ばれます．これらのクラスを特定して理解することに関心があります．



2. 推定

潜在クラスモデルをフィットさせるには，潜在変数にクラスの数を指定する必要があります．ここで示した潜在クラスモデルの基本的な形式では，2 つのクラスを持つ 1 つのカテゴリ潜在変数があります．モデルのパラメータ，つまり，観測された 4 つの変数のロジスティック回帰モデルの切片は，クラス間で異なることが許可されています．

ここで使用するデータセットは Stouffer and Toby (1951)を参考にしています．変数は，4 つの状況にどのように反応するかを尋ねられたハーバード大学とラドクリフ大学の学生の回答を表しています．回答者は，特定の反応（友人への義務に基づく）または普遍的な反応（社会への義務に基づく）のいずれかを選択しました．各変数 [1][2][3][4]は，特定の応答を示す 0 と普遍的な応答を示す 1 でコード化されます．

まず，下記でサンプルファイルをダウンロードします．`describe` で内容を確認します．

```
webuse gsem_lca1, clear
describe
```

---

[1] accident 変数は，次の質問への回答を記録します：あなたは親しい友人が運転する車に乗っていて，彼は歩行者に衝突しました．彼は時速 20 マイルのスピードゾーンで少なくとも時速 35 マイルで走っていました．他に目撃者はいません．彼の弁護士は，速度が時速 20 マイルにすぎなかったことを宣誓の下で証言すれば，彼を深刻な結果から救うかもしれないと言っています．宣誓証人の義務とあなたの友人に対する義務を考慮して，あなたはおそらく何をすると思いますか?

[2] play は，次の質問に対する応答を記録します：あなたはニューヨークのドラマ評論家です．あなたの親しい友人が，ブロードウェイの新作で貯金をすべてつぎ込んでしまいました．あなたは本当にその芝居がダメだと思っています．あなたの読者に対するあなたの義務とあなたの友人に対するあなたの義務を考慮して，あなたのレビューで彼の遊びに気楽に行きますか?

[3] insurance は，次の質問に対する応答を記録します：あなたは保険会社の医師です．あなたは，より多くの保険料が必要な親しい友人を診察します．彼はかなり良い状態であることがわかりますが，診断が難しい小さな点が 1 つまたは 2 つあります．保険会社に対するあなたの義務と，あなたの友人に対するあなたの義務を考慮して，彼に有利なように疑いを覆い隠しますか?

[4] stock は，質問に対する回答を記録します：あなたは会社の取締役会の秘密会議から戻ってきました．取締役会の決定が公表される前に彼が市場から撤退しない限り，破産になる親しい友人がいます．あなたはたまたまその友人の家で同じ晩に夕食をとっています．会社に対するあなたの義務と，友人に対する義務を考慮して，彼にこの件を報告しますか?

ここでは，潜在クラスが 2 つあることを仮定し，変数ごとに定数項のみが異なるモデルを想定する場合，次のように入力し，推定を行います．

```
gsem (accident play insurance stock <- ), logit lclass(C 2)
```

矢印（<-）の左側が観測変数のリスト，今回は定数項のみとなるので右側は空欄です．

カンマ以降はオプションです．ここでは観測変数がバイナリなので，logit オプションでロジットモデルを指定します．lclass オプションは潜在クラスの名前と数を指定します．ここでは名前は C として，2 クラス存在していることを想定するので 2 と入力します．

```
Fitting full model:

Iteration 0:   log likelihood = -504.62913
Iteration 1:   log likelihood = -504.47255
Iteration 2:   log likelihood = -504.46773
Iteration 3:   log likelihood = -504.46767
Iteration 4:   log likelihood = -504.46767

Generalized structural equation model                     Number of obs = 216
Log likelihood = -504.46767
```

	Coefficient	Std. err.	z	P>|z|	[95% conf. interval]	
1.C	(base outcome)					
2.C						
_cons	-.9482041	.2886333	-3.29	0.001	-1.513915	-.3824933

Class: 1

Response: accident
Family: Bernoulli
Link: Logit

Response: play
Family: Bernoulli
Link: Logit

Response: insurance
Family: Bernoulli
Link: Logit

Response: stock
Family: Bernoulli
Link: Logit

	Coefficient	Std. err.	z	P>|z|	[95% conf.	interval]
accident						
_cons	.9128742	.1974695	4.62	0.000	.5258411	1.299907
play						
_cons	-.7099072	.2249096	-3.16	0.002	-1.150722	-.2690926
insurance						
_cons	-.6014307	.2123096	-2.83	0.005	-1.01755	-.1853115
stock						
_cons	-1.880142	.3337665	-5.63	0.000	-2.534312	-1.225972

```
Class:    2

Response: accident
Family:   Bernoulli
Link:     Logit

Response: play
Family:   Bernoulli
Link:     Logit

Response: insurance
Family:   Bernoulli
Link:     Logit

Response: stock
Family:   Bernoulli
Link:     Logit
```

	Coefficient	Std. err.	z	P>|z|	[95% conf.	interval]
accident						
_cons	4.983017	3.745987	1.33	0.183	-2.358982	12.32502
play						
_cons	2.747366	1.165853	2.36	0.018	.4623372	5.032395
insurance						
_cons	2.534582	.9644841	2.63	0.009	.6442279	4.424936
stock						
_cons	1.203416	.5361735	2.24	0.025	.1525356	2.254297

推定結果は反復計算のログから始まります. 最初の 3 つの部分は初期値の計算を行っています. 出力結果の最初の表は, C に対する多項ロジットモデルの推定係数を示しています. 次の 2 つの表は, 1 番目と 2 番目のクラスのロジスティック回帰モデルの結果です.

3. 適合度診断とモデル比較

潜在クラス分析では、推定後にモデルの適合度を診断することができます。観測変数がカテゴリカルな標準的な潜在クラス分析では、推定したモデルと飽和モデルを比較することです。`estat lcgof` を使用し、これらの尤度比検定を行います。この尤度比統計量は、潜在クラス分析では、特にG^2と呼ばれることがあります。

```
estat lcgof
```

Fit statistic	Value	Description
Likelihood ratio		
chi2_ms(6)	2.720	model vs. saturated
p > chi2	0.843	
Information criteria		
AIC	1026.935	Akaike's information criterion
BIC	1057.313	Bayesian information criterion

この結果からは、推定したモデルが飽和モデルと同様に適合するという帰無仮説を棄却できません。

異なるクラス数を想定したモデル同士の適合度を比較することも可能です。まずは，すでに推定した 2 クラスモデルを保存します。

```
estimates store twoclass
```

次に 1 クラスモデルを推定・結果を保存し、飽和モデルと尤度比検定を行います。

```
gsem (accident play insurance stock <- ), logit lclass(C 1)
estimates store oneclass
estat lcgof
```

Fit statistic	Value	Description
Likelihood ratio		
chi2_ms(11)	81.084	model vs. saturated
p > chi2	0.000	
Information criteria		
AIC	1095.300	Akaike's information criterion
BIC	1108.801	Bayesian information criterion

1 クラスモデルでは帰無仮説が棄却され、当てはまりが良くないことがわかります。

さらに 3 クラスモデルを推定し、適合度を確認します。

```
gsem (accident play insurance stock <- ), logit lclass(C 3)
estimates store threeclass
estat lcgof
```

Fit statistic	Value	Description
Likelihood ratio		
chi2_ms(1)	0.387	model vs. saturated
p > chi2	0.534	
Information criteria		
AIC	1034.602	Akaike's information criterion
BIC	1081.856	Bayesian information criterion

この結果からは 3 クラスモデルの適合度は、2 クラスモデルと同様に、飽和モデルと比較しても悪くありません。

保存した 2 クラス、1 クラス、3 クラスモデルを estimates stats コマンドで AIC、BIC を基準にして比較します。

```
estimates stats oneclass twoclass threeclass
```

```
Akaike's information criterion and Bayesian information criterion
```

Model	N	ll(null)	ll(model)	df	AIC	BIC
oneclass	216	.	-543.6498	4	1095.3	1108.801
twoclass	216	.	-504.4677	9	1026.935	1057.313
threeclass	216	.	-503.3011	14	1034.602	1081.856

```
Note: BIC uses N = number of observations. See [R] BIC note.
```

根結果からは、2 クラスモデルがどちらの情報規準について最小の値を持つことがわかります。

4. 予測

このモデルをよりよく理解するために，質問への回答の確率がクラス間でどのように異なるかを調べてみましょう.まず、2 クラスモデルを再推定します。estat lcmean コマンドは，変数ごとにクラス固有の限界平均を報告します.

```
gsem (accident play insurance stock <- ), logit lclass(C 2)
estat lcmean
```

```
Latent class marginal means                                   Number of obs = 216
```

	Delta-method			
	Margin	std. err.	[95% conf.	interval]
1				
accident	.7135879	.0403588	.6285126	.7858194
play	.3296193	.0496984	.2403573	.4331299
insurance	.3540164	.0485528	.2655049	.4538042
stock	.1323726	.0383331	.0734875	.2268872
2				
accident	.9931933	.0253243	.0863544	.9999956
play	.9397644	.0659957	.6135685	.9935191
insurance	.9265309	.0656538	.6557086	.9881667
stock	.769132	.0952072	.5380601	.9050206

この表の最初のセクションは，クラス 1 の確率を報告します. 最初の表から，クラス 1 では, accident 変数に対して 1 と回答をする確率は 0.714 です. 表の 2 番目のセクションも同様に，クラス 2 の対応する確率を報告します.

`estat lcprob` を使用して，各クラスに属する確率を推定できます.

```
estat lcprob
```

```
Latent class marginal probabilities                           Number of obs = 216
```

	Delta-method			
	Margin	std. err.	[95% conf.	interval]
C				
1	.7207539	.0580926	.5944743	.8196407
2	.2792461	.0580926	.1803593	.4055257

これは，データセットに含まれる標本の 72%がクラス 1 に属し, 28%がクラス 2 に属すると予想されることを示しています.

`predict` コマンドでは，クラスメンバーシップの事後確率の予測を使用して，各観測値がどのクラスに属する確率を評価できます. 標本ごと，クラスごとの確率を計算し，1 行目の観測値の確率を `list` コマンドで表示させます.

```
predict classpost*, classposteriorpr
list in 1
```

新しく作成された変数 classpost1 および classpost2 に各クラスとなる確率が保存されます．この新しい変数名の接頭辞は分析者が定義します（ここでは classpost）．アスタリスクは各潜在クラスの番号に置き換えられます.

この推定結果から，各観測値を, 0.5 を基準として，想定した潜在クラスに分類する際には，次のように入力し，新しいカテゴリカル変数を作成します.

```
generate expclass = 1 + (classpost2>0.5)
```

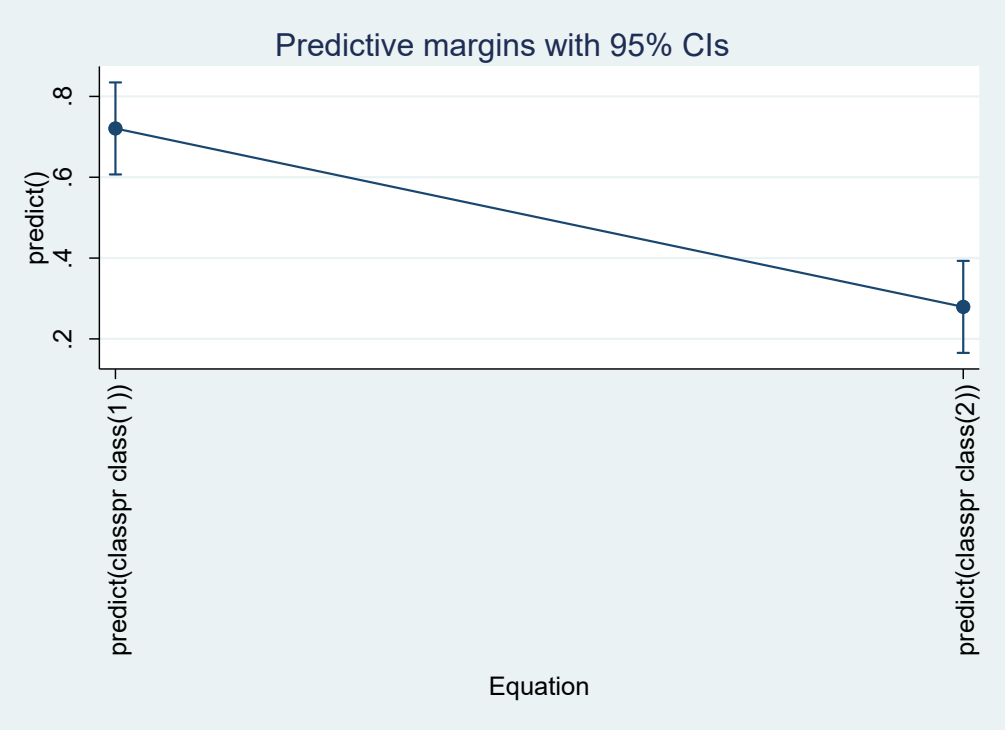
できましたら, tabulate コマンドで構成を確認できます.

```
tabulate expclass
```

expclass	Freq.	Percent	Cum.
1	145	67.13	67.13
2	71	32.87	100.00
Total	216	100.00	

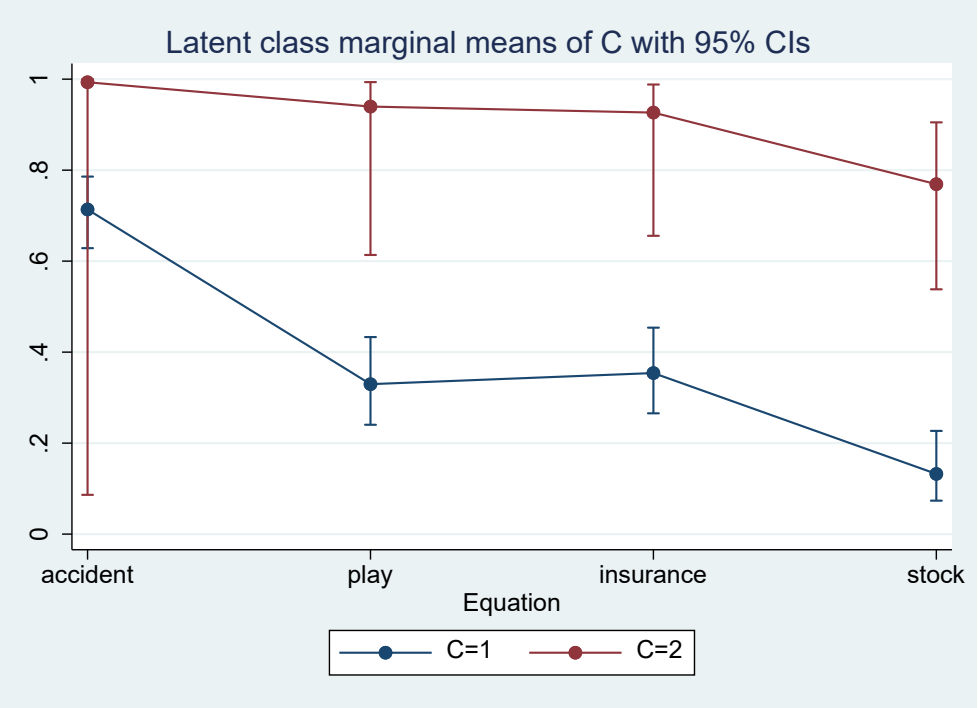
margins と marginsplot を組み合わせて、クラスごと、変数ごとの限界確率を予測・作図できます。まずは、全体的な潜在クラスの限界確率を信頼区間込みで計算します。

```
margins, predict(classpr class(1)) predict(classpr class(2))
marginsplot, xlabel(, angle(vertical))
```



クラスごと・観測変数ごとの限界効果は前述の `estat lcmean` と `marginsplot` で作図できます。

```
estat lcmean
marginsplot
```



5. 潜在プロファイル分析

潜在クラス分析では、「2. 推定」で紹介したように観測変数がカテゴリカルなモデルをフィットしますが、観測変数は離散値に限りません。連続値である場合は、特に潜在プロファイル分析と呼ばれます。

サンプルデータセットをダウンロードし，内容を確認します。

```
use https://www.stata-press.com/data/r17/gsem_lca2, clear
describe
```

Masyn (2013)は、上記のデータを使用して一連の潜在プロファイルモデルに適合させ、それぞれが 1 つのカテゴリ潜在変数と 3 つの観測変数（glucose、insulin、sspg）を持ちます。目標は、これら 3 つの変数に基づいて糖尿病のカテゴリを決定することです。まず、潜在変数 C が 2 つのクラスを持つモデルを当てはめます。

各観測変数ごとに線形回帰モデルを推定します。切片α_{jc}は潜在変数のクラス間で異なることを許可します。さらに、各モデルの誤差分散も推定します。

クラス 1 では次を推定します。

$$glucose = \alpha_{11} + e.glucose$$
$$insulin = \alpha_{21} + e.insulin$$
$$sspg = \alpha_{31} + e.sspg$$

クラス 2 では次を推定します。

$$glucose = \alpha_{12} + e.glucose$$
$$insulin = \alpha_{22} + e.insulin$$
$$sspg = \alpha_{32} + e.sspg$$

さらに、多項ロジット回帰を利用して、各クラスに属する確率を推定します。

$$\Pr(C = 1) = \frac{e^{\gamma_1}}{e^{\gamma_1} + e^{\gamma_2}}$$
$$\Pr(C = 2) = \frac{e^{\gamma_2}}{e^{\gamma_1} + e^{\gamma_2}}$$

γ_1とγ_2は多項ロジットモデルの切片です。デフォルトで、クラス 1 を基準として扱うので、$\gamma_1 = 0$です。デフォルトで誤差は相関せず、分散はクラス間で異なることはないと仮定します。推定コマンドは潜在クラス分析と同様です。

```
gsem (glucose insulin sspg <- _cons), lclass(C 2)
```

```
Generalized structural equation model                     Number of obs = 145
Log likelihood = -1702.5542

 ( 1)  [/]var(e.glucose)#1bn.C - [/]var(e.glucose)#2.C = 0
 ( 2)  [/]var(e.insulin)#1bn.C - [/]var(e.insulin)#2.C = 0
 ( 3)  [/]var(e.sspg)#1bn.C - [/]var(e.sspg)#2.C = 0
```

	Coefficient	Std. err.	z	P>|z|	[95% conf.	interval]
1.C	(base outcome)					
2.C						
_cons	-1.541025	.2205682	-6.99	0.000	-1.973331	-1.10872

```
Class:    1

Response: glucose
Family:   Gaussian
Link:     Identity

Response: insulin
Family:   Gaussian
Link:     Identity

Response: sspg
Family:   Gaussian
Link:     Identity
```

	Coefficient	Std. err.	z	P>|z|	[95% conf.	interval]
glucose						
_cons	41.22237	1.298051	31.76	0.000	38.67824	43.7665
insulin						
_cons	20.98005	1.000974	20.96	0.000	19.01817	22.94192
sspg						
_cons	14.96579	.6868081	21.79	0.000	13.61967	16.31191
var(e.glucose)	191.5596	23.83815			150.0992	244.4723
var(e.insulin)	119.0542	14.00336			94.54204	149.9217
var(e.sspg)	55.91283	6.713667			44.18801	70.7487

```
Class:    2

Response: glucose
Family:   Gaussian
Link:     Identity

Response: insulin
Family:   Gaussian
Link:     Identity

Response: sspg
Family:   Gaussian
Link:     Identity
```

	Coefficient	Std. err.	z	P>|z|	[95% conf.	interval]
glucose						
_cons	115.7123	2.849914	40.60	0.000	110.1266	121.2981
insulin						
_cons	7.553144	2.160949	3.50	0.000	3.317761	11.78853
sspg						
_cons	34.5529	1.53117	22.57	0.000	31.55187	37.55394
var(e.glucose)	191.5596	23.83815			150.0992	244.4723
var(e.insulin)	119.0542	14.00336			94.54204	149.9217
var(e.sspg)	55.91283	6.713667			44.18801	70.7487

最初の表には潜在クラス C の多項ロジットモデルで推定された係数が報告されます。次の 2 つの表は、クラスごとの線形回帰モデルの結果を報告します。

推定後は、潜在クラスモデルと同様に、複数の異なるモデルを比較することが可能です。

5.1 共分散のある 3 クラスモデル

Masyn (2013)の最終的なモデルは、誤差項間の共分散を考慮し、すべてのパラメータをクラス間で個別に推定する 3 クラスモデルでした。共分散を推定するために、covstructure(e._0En, unstructed)を追加します。また、クラス間で全てのパラメータが異なることを許容するため、lcinvariant(none)オプションを追加します。

```
gsem (glucose insulin sspg <- _cons), lclass(C 3) lcinvariant(none)
covstructure(e._OEn, unstructured)
```

```
Generalized structural equation model                     Number of obs = 145
Log likelihood = -1536.6409
```

	Coefficient	Std. err.	z	P>|z|	[95% conf.	interval]
1.C	(base outcome)					
2.C						
_cons	-.8853513	.2386536	-3.71	0.000	-1.353104	-.4175988
3.C						
_cons	-.612664	.2260018	-2.71	0.007	-1.055619	-.1697085

```
Class:    1

Response: glucose
Family:   Gaussian
Link:     Identity

Response: insulin
Family:   Gaussian
Link:     Identity

Response: sspg
Family:   Gaussian
Link:     Identity
```

	Coefficient	Std. err.	z	P>|z|	[95% conf.	interval]
glucose						
_cons	35.68584	.5741752	62.15	0.000	34.56048	36.81121
insulin						
_cons	16.58066	.6204724	26.72	0.000	15.36456	17.79677
sspg						
_cons	10.49755	.5833606	17.99	0.000	9.354183	11.64091
var(e.glucose)	19.30952	3.932547			12.9544	28.78233
var(e.insulin)	26.7354	4.494093			19.23108	37.16804
var(e.sspg)	18.71079	3.970509			12.34422	28.36094
cov(e.glucose,e.insulin)	3.456027	2.942391	1.17	0.240	-2.310954	9.223008
cov(e.glucose,e.sspg)	5.474303	2.811729	1.95	0.052	-.0365846	10.98519
cov(e.insulin,e.sspg)	7.995803	3.020304	2.65	0.008	2.076115	13.91549

```
Class:    2

Response: glucose
Family:   Gaussian
Link:     Identity

Response: insulin
Family:   Gaussian
Link:     Identity

Response: sspg
Family:   Gaussian
Link:     Identity
```

	Coefficient	Std. err.	z	P>|z|	[95% conf.	interval]
glucose						
_cons	47.66176	1.492718	31.93	0.000	44.73609	50.58744
insulin						
_cons	34.35203	3.00337	11.44	0.000	28.46554	40.23853
sspg						
_cons	24.414	.7395383	33.01	0.000	22.96453	25.86347
var(e.glucose)	53.21326	15.56547			29.99396	94.40735
var(e.insulin)	228.6332	59.03553			137.832	379.2526
var(e.sspg)	13.75515	3.838523			7.960284	23.76853
cov(e.glucose,e.insulin)	40.02875	23.12762	1.73	0.083	-5.300552	85.35805
cov(e.glucose,e.sspg)	.7294854	5.48065	0.13	0.894	-10.01239	11.47136
cov(e.insulin,e.sspg)	-5.743169	11.4943	-0.50	0.617	-28.27158	16.78524

```
Class:    3

Response: glucose
Family:   Gaussian
Link:     Identity

Response: insulin
Family:   Gaussian
Link:     Identity

Response: sspg
Family:   Gaussian
Link:     Identity
```

	Coefficient	Std. err.	z	P>|z|	[95% conf.	interval]
glucose						
_cons	93.92473	6.985336	13.45	0.000	80.23372	107.6157
insulin						
_cons	10.37614	1.123135	9.24	0.000	8.174836	12.57744
sspg						
_cons	28.4787	1.94975	14.61	0.000	24.65726	32.30013
var(e.glucose)	1279.011	312.6774			792.1048	2065.218
var(e.insulin)	36.38521	9.26287			22.09163	59.92692
var(e.sspg)	113.3239	27.67628			70.21642	182.8961
cov(e.glucose,e.insulin)	-163.4383	47.637	-3.43	0.001	-256.8051	-70.07153
cov(e.glucose,e.sspg)	276.9206	81.60543	3.39	0.001	116.9769	436.8643
cov(e.insulin,e.sspg)	-25.4313	11.66564	-2.18	0.029	-48.29554	-2.567057

このモデルでは独立変数が無いため、切片がそれぞれ対応する変数の、クラスごとの平均値を表しています。クラス 1 では、glucose の予測された平均値は 35.69, insulin は 16.58, sspg は 10.50 です。分散と共分散についても同様に、各変数のクラスごとの分散と共分散を表します。クラス 1 では、glucose の分散は 19.31, glucose と insulin の共分散は 3.46 となっています。

推定後には、潜在クラスモデルと同様に predict コマンドで事後クラス確率から各標本の分類予測が可能です。

```
predict cpost*, classposteriorpr
egen max = rowmax(cpost*)
generate predclass = 1 if cpost1==max
replace predclass = 2 if cpost2==max
replace predclass = 3 if cpost3==max
tabulate cclass predclass, col
```

Key
frequency *column percentage*

Clinical classification	predclass 1	2	3	Total
Overt diabetic	0	2	31	33
	0.00	6.25	83.78	22.76
Chemical diabetic	7	23	6	36
	9.21	71.88	16.22	24.83
Normal	69	7	0	76
	90.79	21.88	0.00	52.41
Total	76	32	37	145
	100.00	100.00	100.00	100.00

参考文献

Masyn, K. E. 2013. Latent class analysis and finite mixture modeling. In The Oxford Handbook of Quantitative Methods, ed. T. D. Little, vol. 2, 551–610. New York: Oxford University Press.

Samuel A. Stouffer and Jackson Toby. 1951. Role conflict and personality. *American Journal of Sociology* 56: 395-406.