Generalized structural equation models: GLMs, multilevel, latent classes, and more

Kristin MacDonald

Executive Director of Statistical Services
StataCorp LLC

Italian Stata Conference

Outline:

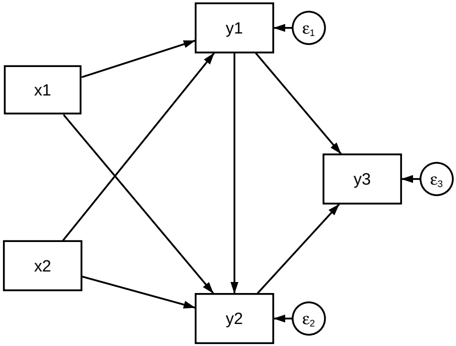
- Introduction to generalized structural equation models
- Introduction to the `gsem` command
- Examples
    - Path models
    - Confirmatory factor models
    - Structural equation models
    - Latent class models
    - Extensions

Review of standard structural equation models

- What are structural equation models?
  - Structural equation models encompass a wide range of models.
  - They allow for fitting multiple equations simultaneously.
  - They may incorporate both observed and latent variables.

- The `sem` command fits one-level, linear structural equation models.
- Some common examples of these models include path models, confirmatory factor models, and full structural equation models.
- These models are often described in path diagrams.

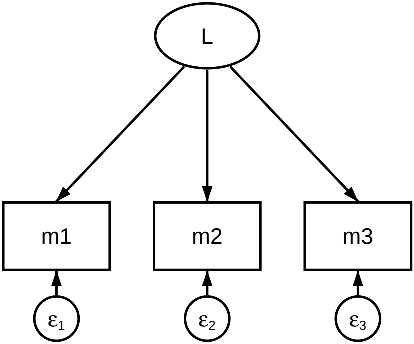
- Models with multiple equations that model relationships among observed variables are often called path models.
- For example, a path diagram may be



- To fit this model we type

```
. sem (y1 <- x1 x2) ///
      (y2 <- y1 x1 x2) ///
      (y3 <- y1 y2)
```

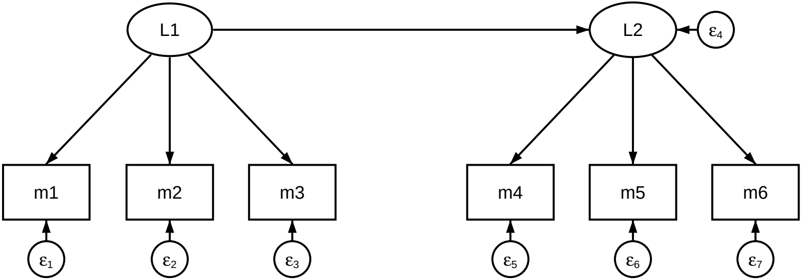
- Models of relationships between latent variables and the observed variables that measure the latent concepts are often referred to as confirmatory factor models or measurement models.
- For example, a path diagram may be



- To fit this model we type

```
.   sem (L -> m1 m2 m3)
```

- Full structural equation models combine elements of path models and confirmatory factor models, having both measurement components and structural components.
- For example, a path diagram may be



- To fit this model we type

```
. sem (L1 -> m1 m2 m3) ///
      (L2 -> m4 m5 m6) ///
      (L2 <- L1)
```

Generalized structural equation models

- What are generalized structural equation models?
- These models extend the standard linear structural equation models in three ways:
    1. Generalized linear models
    2. Categorical latent variables
    3. Multilevel models
- The `gsem` command fits these models.

Introduction to gsem

- The basic syntax for `gsem` is

  `gsem` *paths* `[if]` `[in]` *weight*, *options*
- *paths* specify direct paths between variables

  `(`*names1* `<-` *names2*`)`

  or

  `(`*names2* `->` *names1*`)`
- *names1* and *names2* may contain lists of observed and latent variables. By default, lowercase variables are observed and capitalized variables are latent.
- For example, for linear regression, type

```
. gsem (y <- x1 x2)
```

- Options are available to specify generalized linear models.
- The `family()` and `link()` options may be specified for individual paths or globally.
- The default is the Gaussian family and identity link.

	identity	log	logit	probit	cloglog
gaussian	D	x			
bernoulli			D	x	x
beta			D	x	x
binomial			D	x	x
ordinal			D	x	x
multinomial			D		
poisson		D			
nbinomial		D			
exponential		D			
weibull		D			
gamma		D			
loglogistic		D			
lognormal		D			
pointmass	D				

Option	Synonym for
exponential	synonym for family(exponential) link(log)
gamma	synonym for family(gamma) link(log)
logit	synonym for family(bernoulli) link(logit)
loglogistic	synonym for family(loglogistic) link(log)
lognormal	synonym for family(lognormal) link(log)
llogistic	synonym for family(llogistic) link(log)
lnormal	synonym for family(lnormal) link(log)
mlogit	synonym for family(multinomial) link(logit)
nbreg	synonym for family(nbinomial mean) link(log)
ocloglog	synonym for family(ordinal) link(cloglog)
ologit	synonym for family(ordinal) link(logit)
oprobit	synonym for family(ordinal) link(probit)
poisson	synonym for family(poisson) link(log)
probit	synonym for family(bernoulli) link(probit)
regress	synonym for family(gaussian) link(identity)
weibull	synonym for family(weibull) link(log)

- For example, to fit a logistic regression, we one of

```
.   gsem (y <- x1 x2,
          family(bernoulli) link(logit))
.   gsem (y <- x1 x2),
          family(bernoulli) link(logit)
.   gsem (y <- x1 x2, logit)
.   gsem (y <- x1 x2), logit
```

- Observation-level continuous latent variables are specified by capitalization or via the `latent()` option.

```
. gsem (L -> y1 y2 y3)
```

or

```
. gsem (l -> y1 y2 y3), latent(l)
```

- Observation-level categorical latent variables are specified via the `lclass()` option. The number of categories is also specified.

```
.   gsem (y1 y2 y3 <- ), lclass(C 2)
```

- Multilevel models are fit by specifying the level at which a latent variable varies in square brackets.
- For data with children nested in schools, to fit a two-level random-intercept model, we could include latent variable `RI` that varies at the school level

```
. gsem (y <- x1 x2 RI[school])
```

- We can include a random slope by interacting a level-two latent variable with an observed variable.

```
. gsem (y <- x1 x2 RI[school] c.x1#RS[school])
```

- To fit a three-level or higher-level model, specify the nesting structure in square brackets.

```
. gsem (y <- x1 x2 Rsch[school]
        Rcl[school>class])
```

Examples

Example 1: Path model (mediation model)

- We can fit path models with `gsem` similarly to fitting them with `sem`.
- To demonstrate, we use fictional data on job performance, job satisfaction, and perceived support from managers.
- The data represent 1,500 sales employees of a large department store in 75 locations.
- We fit a mediation model where managerial support is expected to affect job performance directly and indirectly via job satisfaction.

```
. webuse gsem_multmed, clear
(Fictional job-performance data)

. describe

Contains data from https://www.stata-press.com/data/r19/gsem_multmed.dta
 Observations:         1,500                  Fictional job-performance data
    Variables:             4                  26 Mar 2024 10:17
                                              (_dta has notes)
--------------------------------------------------------------------------------
Variable      Storage   Display    Value
    name         type    format    label      Variable label
--------------------------------------------------------------------------------
branch          byte    %9.0g                 Department store branch
support         float   %9.0g                 Perceived supervisor support
satis           float   %9.0g                 Job satisfaction
perform         float   %9.0g                 Job performance
--------------------------------------------------------------------------------
Sorted by:
```

- We can fit the linear model with the `sem` command.

```
.   sem (perform <- satis support) ///
        (satis <- support)
```

- We can fit the same model with `gsem`.

```
. gsem (perform <- satis support) (satis <- support)

Iteration 0:  Log likelihood = -2674.3421
Iteration 1:  Log likelihood = -2674.3421

Generalized structural equation model                Number of obs = 1,500

Response: perform
Family:   Gaussian
Link:     Identity

Response: satis
Family:   Gaussian
Link:     Identity

Log likelihood = -2674.3421
```

	Coefficient	Std. err.	z	P>|z|	[95% conf.	interval]
perform						
satis	.8984401	.0251903	35.67	0.000	.849068	.9478123
support	.6161077	.0303143	20.32	0.000	.5566927	.6755227
_cons	4.981054	.0150589	330.77	0.000	4.951539	5.010569
satis						
support	.2288945	.0305047	7.50	0.000	.1691064	.2886826
_cons	.019262	.0154273	1.25	0.212	-.0109749	.0494989
var(e.perform)	.3397087	.0124044			.3162461	.364912
var(e.satis)	.3569007	.0130322			.3322507	.3833795

- With gsem, we can fit a multilevel model that accounts for employees being nested within branches.
- We include random intercepts in the models for both perform and satis.
- Then we use nlcom to compute indirect and total effects of managerial support on performance.

```
. gsem (perform <- satis support M1[branch]) ///
>       (satis <- support M2[branch]), ///
>       cov(M1[branch]*M2[branch]@0) nolog noheader
 ( 1)  [perform]M1[branch] = 1
 ( 2)  [satis]M2[branch] = 1
```

	Coefficient	Std. err.	z	P>|z|	[95% conf.	interval]
perform						
satis	.604264	.0336398	17.96	0.000	.5383313	.6701968
support	.6981525	.0250432	27.88	0.000	.6490687	.7472364
M1[branch]	1	(constrained)				
_cons	4.986596	.0489465	101.88	0.000	4.890663	5.082529
satis						
support	.2692633	.0179649	14.99	0.000	.2340528	.3044739
M2[branch]	1	(constrained)				
_cons	.0189202	.0570868	0.33	0.740	-.0929678	.1308083
var(M1[bra~h])	.1695962	.0302866			.119511	.2406713
var(M2[bra~h])	.2384738	.0399154			.1717781	.3310652
var(e.perform)	.201053	.0075451			.1867957	.2163985
var(e.satis)	.1188436	.0044523			.1104299	.1278983

- Indirect effect of support on performance:

```
. nlcom _b[perform:satis]*_b[satis:support]
       _nl_1: _b[perform:satis]*_b[satis:support]
```

	Coefficient	Std. err.	z	P>|z|	[95% conf.	interval]
_nl_1	.1627062	.0141382	11.51	0.000	.1349958	.1904165

- Total effect of support on performance:

```
. nlcom _b[perform:support]+_b[perform:satis]*_b[satis:support]
       _nl_1: _b[perform:support]+_b[perform:satis]*_b[satis:support]
```

	Coefficient	Std. err.	z	P>|z|	[95% conf.	interval]
_nl_1	.8608587	.0257501	33.43	0.000	.8103894	.911328

Example 2: Confirmatory factor model (item response theory model)

- To demonstrate a confirmatory factor model, we have fictional data on 500 students and their responses to five questions designed to measure math ability.
- Variables q1 through q5 are recorded as 1 if the question is answered correctly and 0 if answered incorrectly.

```
. use math, clear
(Fictional math abilities data)

. describe id q1-q5

Variable      Storage   Display    Value
    name         type    format    label      Variable label
-------------------------------------------------------------------------------
id               long    %9.0g                Student id
q1               byte    %9.0g     result     q1 correct
q2               byte    %9.0g     result     q2 correct
q3               byte    %9.0g     result     q3 correct
q4               byte    %9.0g     result     q4 correct
q5               byte    %9.0g     result     q5 correct
```

- To fit a logit model for each of the five outcomes, we specify the `logit` option.
- We must have constraints to give the latent variable a location and scale. By default, the mean is constrained to 0 and the variance is set by constraining the coefficient on the latent variable in the first equation to 1. Instead, we specify `variance(MathAb@1)` to set the variance to 1 so that we can easily compare coefficients.

```
. gsem (MathAb -> q1-q5), logit variance(MathAb@1) nolog noheader
 ( 1)  [/]var(MathAb) = 1
```

	Coefficient	Std. err.	z	P>|z|	[95% conf.	interval]
q1						
MathAb	1.521243	.4062756	3.74	0.000	.7249571	2.317528
_cons	.0360205	.1274819	0.28	0.778	-.2138394	.2858804
q2						
MathAb	.3876489	.142974	2.71	0.007	.1074251	.6678727
_cons	-.4458798	.0953314	-4.68	0.000	-.6327259	-.2590336
q3						
MathAb	.7244121	.1715327	4.22	0.000	.3882142	1.06061
_cons	.1525973	.1004485	1.52	0.129	-.044278	.3494727
q4						
MathAb	.4074442	.1428405	2.85	0.004	.1274819	.6874066
_cons	-.3184607	.094383	-3.37	0.001	-.5034479	-.1334734
q5						
MathAb	1.458619	.3783345	3.86	0.000	.7170966	2.200141
_cons	-.0541335	.125028	-0.43	0.665	-.2991838	.1909168
var(MathAb)	1	(constrained)				

- The coefficient on `MathAb` is positive in in all 5 equations, indicating that as `MathAb` increases, the probability of answering the questions correctly also increases.
- The coefficient is largest for question `q1`. In item response theory, we say that this question has the highest discrimination (or ability to separate lower levels of math ability from higher levels).
- For standard item response theory models, use the `irt` suite of commands.

Example 3: Full structural equation model

- We extend the previous model with a second measurement component and a structural component that regresses one latent variable on the other.
- Variables `att1` through `att5` are ordinal variables measuring students' attitudes toward math.

```
. describe att1-att5

Variable      Storage   Display    Value
    name         type    format    label      Variable label
-------------------------------------------------------------------------------
att1            float    %26.0g    agree      Skills taught in math class will
                                                help me get a better job.
att2            float    %26.0g    agree      Math is important in everyday life
att3            float    %26.0g    agree      Working math problems makes me
                                                anxious.
att4            float    %26.0g    agree      Math has always been my worst
                                                subject.
att5            float    %26.0g    agree      I am able to learn new math
                                                concepts easily.
```

- Latent variable `MathAtt` is measured by these variables, and we use ordinal logit models for each one.
- We are interested in the effect of `MathAtt` on `MathAb`, so we include this path in the model.

```
. gsem (MathAb -> q1-q5, logit) ///
>        (MathAtt -> att1-att5, ologit) ///
>        (MathAtt -> MathAb), nolog noheader
 ( 1)   [q1]MathAb = 1
 ( 2)   [att1]MathAtt = 1
```

	Coefficient	Std. err.	z	P>|z|	[95% conf.	interval]
q1						
MathAb	1	(constrained)				
_cons	.0403127	.1274162	0.32	0.752	-.2094184	.2900439
q2						
MathAb	.2393296	.1007691	2.38	0.018	.0418258	.4368333
_cons	-.4440017	.0948264	-4.68	0.000	-.6298579	-.2581454
q3						
MathAb	.5884097	.1950271	3.02	0.003	.2061636	.9706558
_cons	.1615205	.1054903	1.53	0.126	-.0452366	.3682776
q4						
MathAb	.2250662	.0973337	2.31	0.021	.0342957	.4158368
_cons	-.3149013	.0932356	-3.38	0.001	-.4976396	-.1321629
q5						
MathAb	.8568504	.263716	3.25	0.001	.3399765	1.373724
_cons	-.0486802	.1188913	-0.41	0.682	-.2817029	.1843426

att1						
MathAtt	1	(constrained)				
att2						
MathAtt	.3754426	.0953942	3.94	0.000	.1884733	.5624118
att3						
MathAtt	-1.468288	.3242252	-4.53	0.000	-2.103757	-.8328178
att4						
MathAtt	-.7841452	.1496076	-5.24	0.000	-1.077371	-.4909197
att5						
MathAtt	.5079287	.1147362	4.43	0.000	.2830498	.7328075
MathAb						
MathAtt	.622715	.1680986	3.70	0.000	.2932478	.9521823

/att1					
cut1	-1.116744	.1336836		-1.378759	-.8547286
cut2	-.2504544	.1176572		-.4810582	-.0198506
cut3	.3058712	.1181599		.0742821	.5374602
cut4	1.352897	.1420803		1.074424	1.631369
/att2					
cut1	-1.05734	.106547		-1.266168	-.8485113
cut2	-.1938871	.094358		-.3788254	-.0089489
cut3	.3608793	.0954384		.1738236	.5479351
cut4	1.135196	.1085201		.9225009	1.347892
/att3					
cut1	-1.02814	.1655276		-1.352569	-.7037123
cut2	-.0495702	.1404696		-.3248855	.2257451
cut3	.5417343	.1485789		.2505249	.8329436
cut4	1.621894	.2003441		1.229227	2.014561
/att4					
cut1	-1.073436	.1213559		-1.31129	-.8355832
cut2	-.2106145	.1077409		-.4217829	.0005538
cut3	.407913	.1095998		.1931015	.6227246
cut4	1.401051	.1313979		1.143516	1.658586
/att5					
cut1	-1.24666	.1150748		-1.472203	-1.021118
cut2	-.3384903	.0988109		-.5321561	-.1448246
cut3	.2120879	.0979818		.0200471	.4041288
cut4	.9283652	.1072727		.7181145	1.138616

- The coefficient on `MathAtt` in the `MathAb` equation is 0.62, indicating a positive effect of attitude toward math on abilities.

Example 4: Latent class analysis

- We are interested in finding and understanding groups in a population. These unobserved groups can be represented by a categorical latent variable.
- To demonstrate we use a dataset with students' responses to four questions regarding situations where they would have to choose between obligations to friends versus obligations to society.

```
. webuse gsem_lca1, clear
(Latent class analysis)

. describe

Contains data from https://www.stata-press.com/data/r19/gsem_lca1.dta
 Observations:           216                  Latent class analysis
    Variables:             4                  17 Jan 2025 12:52
                                              (_dta has notes)
-----------------------------------------------------------------------------------
Variable      Storage   Display    Value
    name         type    format    label      Variable label
-----------------------------------------------------------------------------------
accident        byte    %9.0g                 Would testify against friend in
                                                accident case
play            byte    %9.0g                 Would give negative review of
                                                friend's play
insurance       byte    %9.0g                 Would disclose health concerns to
                                                friend's insurance company
stock           byte    %9.0g                 Would keep company secret from friend
-----------------------------------------------------------------------------------
Sorted by: accident  play  insurance  stock
```

```
. notes in 1/4

_dta:
  1.  Source: Data from Stouffer, S. A., and J. Toby. 1951. Role conflict and
      personality. American Journal of Sociology 56: 395-406.
      https://doi.org/10.1086/220785.
  2.  Variables represent responses of students from Harvard and Radcliffe who were
      asked how they would respond to four situations. Respondents selected either
      a particularistic response (based on obligations to a friend) or
      universalistic response (based on obligations to society).
  3.  Each variable is coded with 0 indicating a particularistic response and 1
      indicating a universalistic response.
  4.  For a full description of the questions, type "notes in 5/8".
```

- We first fit a two-class model, allowing intercepts in logit models for each of our variables of interest to vary across the two classes.
- We store the results of this model as `twoclass`.

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

	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

	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

------------------------------------------------------------------------------
             | 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
------------------------------------------------------------------------------

. estimates store twoclass
```

- We also fit a three-class model so that we can evaluate whether two or three classes fit better.
- We store the results of this model as `threeclass`.
- The `lcstats` command provides a variety of statistics that can be used to select from these models.

```
. quietly gsem (accident play insurance stock <- ), logit lclass(C 3)
. estimates store threeclass
. lcstats twoclass threeclass, all split
```

Latent class statistics

	N	Rank	AIC	BIC	AICc	CAIC	Entropy
twoclass	216	9	1,026.94	1,057.31	1,027.81	1,066.31	0.7193
threeclass	216	14	1,034.60	1,081.86	1,036.69	1,095.86	0.6110

AIC is the Akaike information criterion.
BIC is the Bayesian information criterion.
AICc is the corrected Akaike information criterion.
CAIC is the consistent Akaike information criterion.
BIC, AICc, and CAIC use N = number of observations.

	Classes	ll	df	VLMR	P>VLMR	LMR	P>LMR
twoclass	2	-504.47					
threeclass	3	-503.30	5	2.33	0.673	2.25	0.687

VLMR is the Vuong-Lo-Mendell-Rubin likelihood-ratio test statistic.
LMR is the Lo-Mendell-Rubin-adjusted likelihood-ratio test statistic.
Likelihood-ratio tests compare the given model versus the same model with one less latent class.

- All statistics suggest that we select the two-class model.
- We `estimates restore` the results, and the evaluate this model further.
- `estat lcprob` reports the expected proportions in each class.

```
. estimates restore twoclass
(results twoclass are active now)

. estat lcprob

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

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

- We expect 72% in class 1 and 28% in class 2.
- `estat lcmean` allows us to learn more about these two classes.

```
. estat lcmean

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

	Margin	Delta-method 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

- In the first class, the probability of choosing the response that favors society was high (0.71) for the first question, but lower for the others.
- In the second class, the probability of choosing the response that favors society was higher for all four questions.
- Class 2 seems to be more inclined to favor society as a whole.

Example 5: Multilevel-model group comparisons

- When there are known groups in the data, the `group()` option makes it easy to explore group differences.
- Here we fit a multilevel linear model and evaluate whether parameters vary across groups.
- We use data on children's weights from an example in **[ME] mixed**.

```
. webuse childweight
(Weight data on Asian children)

. describe

Contains data from https://www.stata-press.com/data/r19/childweight.dta
 Observations:           198                  Weight data on Asian children
    Variables:             5                  23 May 2024 15:12
                                              (_dta has notes)
-------------------------------------------------------------------------------------
Variable      Storage   Display    Value
    name         type    format    label      Variable label
-------------------------------------------------------------------------------------
id              int     %8.0g                 Child identifier
age             float   %8.0g                 Age in years
weight          float   %8.0g                 Weight in Kg
brthwt          int     %8.0g                 Birthweight in g
girl            byte    %9.0g      bg         Gender
-------------------------------------------------------------------------------------
Sorted by: id  age

. notes

_dta:
  1.  Source: Goldstein, H. 1986. Efficient statistical modelling of longitudinal
      data. Annals of Human Biology 13: 129-141.
      https://doi.org/10.1080/03014468600008271.
```

- With `mixed`, we can use factor variables and interactions to allow coefficients and random effects to vary across groups.
- As shown below, we can also use the `residuals(, by())` option to estimate the error variance separately for groups—boys and girls, in this case.

```
. mixed weight age || id:age, residuals(independent, by(girl)) nolog
Mixed-effects ML regression                     Number of obs     =    198
Group variable: id                              Number of groups  =     68
                                                Obs per group:
                                                              min =      1
                                                              avg =    2.9
                                                              max =      5
                                                Wald chi2(1)      = 710.12
Log likelihood = -342.96155                     Prob > chi2       = 0.0000

-------------------------------------------------------------------------------
      weight | Coefficient  Std. err.      z    P>|z|     [95% conf. interval]
-------------+-----------------------------------------------------------------
         age |   3.452597   .1295629    26.65   0.000     3.198658    3.706535
       _cons |   5.097696   .1559364    32.69   0.000     4.792066    5.403326
-------------------------------------------------------------------------------

-------------------------------------------------------------------------------
  Random-effects parameters  |   Estimate   Std. err.     [95% conf. interval]
-----------------------------+-------------------------------------------------
id: Independent              |
                    var(age) |   .2728324   .1498633      .0929698    .8006633
                  var(_cons) |     .24972   .2313903      .0406191    1.535238
-----------------------------+-------------------------------------------------
Residual: Independent,       |
    by girl                  |
                 Boy: var(e) |     1.5956   .2940092      1.111936    2.289647
                Girl: var(e) |   1.105886   .2115579      .7601058    1.608965
-------------------------------------------------------------------------------
LR test vs. linear model: chi2(3) = 26.89                 Prob > chi2 = 0.0000
```

- With `gsem`, we can use the `group()` option to fit a model that allows any parameters to vary across groups.
- The `ginvariant()` option specifies which types of parameters should be constrained across groups.
- Here we estimate different coefficients, variances of the random intercept, and error variances.

```
. gsem (weight <- age RI[id]), group(girl) ginvariant(loading mean) nolog noheader
Group: Boy                                                  Number of obs = 100
```

	Coefficient	Std. err.	z	P>|z|	[95% conf.	interval]
weight						
age	3.50015	.1815822	19.28	0.000	3.144255	3.856044
RI[id]	1	(constrained)				
_cons	5.398298	.2691278	20.06	0.000	4.870818	5.925779
var(RI[id])	.4854647	.2799216			.1568015	1.503021
var(e.weight)	1.922634	.3240432			1.381764	2.675218

```
Group: Girl                                     Number of obs = 98
```

	Coefficient	Std. err.	z	P>|z|	[95% conf.	interval]
weight						
age	3.277961	.1375257	23.84	0.000	3.008416	3.547506
RI[id]	1	(constrained)				
_cons	4.918627	.2206876	22.29	0.000	4.486087	5.351167
var(RI[id])	.5036919	.2355279			.2014376	1.259475
var(e.weight)	1.080144	.1926215			.7615309	1.532061

- We can then test for equality of parameters.
- For example, we can test whether the coefficient on age is equal for boys and girls.

```
. test _b[weight:0.girl#c.age]=_b[weight:1.girl#c.age]
 ( 1)  [weight]0bn.girl#c.age - [weight]1.girl#c.age = 0
           chi2(  1) =    0.95
         Prob > chi2 =    0.3293
```

Example 6: Cross-equation tests

- With the `suest` command, we can combine results from multiple estimation commands and then perform tests.
- With `gsem` we can fit multiple models simultaneously and then perform similar tests. However, with `gsem`, we can do so with a wider array of models than are supported by `suest`.
- Following an example from **[R] suest**, we consider labor market data for siblings on income and whether or not they got a promotion.

```
. webuse income, clear

. describe

Contains data from https://www.stata-press.com/data/r19/income.dta
 Observations:           277
    Variables:             8                  2 Dec 2024 15:41
--------------------------------------------------------------------------------
Variable      Storage   Display    Value
    name         type    format    label      Variable label
--------------------------------------------------------------------------------
famid           int     %9.0g                 Family ID
e               float   %9.0g
male            byte    %9.0g                 1 if male
edu             float   %9.0g                 Formal education
exp             float   %9.0g                 On-the-job training experience
u               float   %9.0g
inc             float   %9.0g                 Income
promo           byte    %9.0g                 1 if person was promoted last year
--------------------------------------------------------------------------------
Sorted by:
```

- We are interested in determining whether income and getting a promotion are both reflections of the same underlying concept—human capital.
- We fit a linear regression for income, a probit model for getting a promotion, combine the results with `suest`, and then test whether the coefficients are proportional.

```
. regress inc edu exp male

      Source |       SS           df       MS      Number of obs   =       277
-------------+----------------------------------   F(3, 273)       =     42.34
       Model |  2058.44672         3  686.148908   Prob > F        =    0.0000
    Residual |  4424.05183       273  16.2053181   R-squared       =    0.3175
-------------+----------------------------------   Adj R-squared   =    0.3100
       Total |  6482.49855       276  23.4873136   Root MSE        =    4.0256

------------------------------------------------------------------------------
         inc | Coefficient  Std. err.      t    P>|t|     [95% conf. interval]
-------------+----------------------------------------------------------------
         edu |   2.213707    .243247     9.10   0.000     1.734828    2.692585
         exp |    1.47293    .231044     6.38   0.000     1.018076    1.927785
        male |   .5381153   .4949466     1.09   0.278    -.436282     1.512513
       _cons |   1.255497   .3115808     4.03   0.000      .642091    1.868904
------------------------------------------------------------------------------

. est store inc
```

```
. probit promo edu exp male
Iteration 0:   Log likelihood = -183.31675
Iteration 1:   Log likelihood = -158.55292
Iteration 2:   Log likelihood = -158.43891
Iteration 3:   Log likelihood = -158.43888
Iteration 4:   Log likelihood = -158.43888
Probit regression                                       Number of obs =    277
                                                        LR chi2(3)    =  49.76
                                                        Prob > chi2   = 0.0000
Log likelihood = -158.43888                             Pseudo R2     = 0.1357

------------------------------------------------------------------------------
       promo | Coefficient  Std. err.      z    P>|z|     [95% conf. interval]
-------------+----------------------------------------------------------------
         edu |   .4593002   .0898537     5.11   0.000     .2831901    .6354102
         exp |   .3593023   .0805774     4.46   0.000     .2013735    .5172312
        male |   .2079983   .1656413     1.26   0.209    -.1166527    .5326494
       _cons |   -.464622   .1088166    -4.27   0.000    -.6778985   -.2513454
------------------------------------------------------------------------------

. est store promo
```

```
. etable, estimates(inc promo) eqrecode(inc=promo)
```

	inc	promo
Formal education	2.214	0.459
	(0.243)	(0.090)
On-the-job training experience	1.473	0.359
	(0.231)	(0.081)
1 if male	0.538	0.208
	(0.495)	(0.166)
Intercept	1.255	-0.465
	(0.312)	(0.109)
Number of observations	277	277

```
. suest inc promo, vce(cluster famid)
```

Simultaneous results for inc, promo   Number of obs = 277

(Std. err. adjusted for 135 clusters in famid)

	Coefficient	Robust std. err.	z	P>|z|	[95% conf.	interval]
inc_mean						
edu	2.213707	.2483907	8.91	0.000	1.72687	2.700543
exp	1.47293	.1890583	7.79	0.000	1.102383	1.843478
male	.5381153	.4979227	1.08	0.280	-.4377952	1.514026
_cons	1.255497	.3374977	3.72	0.000	.594014	1.916981
inc_lnvar						
_cons	2.785339	.079597	34.99	0.000	2.629332	2.941347
promo_promo						
edu	.4593002	.0886982	5.18	0.000	.2854549	.6331454
exp	.3593023	.079772	4.50	0.000	.2029522	.5156525
male	.2079983	.1691053	1.23	0.219	-.1234419	.5394386
_cons	-.464622	.1042169	-4.46	0.000	-.6688833	-.2603607

```
. testnl [inc_mean]edu/[promo_promo]edu = ///
>        [inc_mean]exp/[promo_promo]exp = ///
>        [inc_mean]male/[promo_promo]male

  (1)  [inc_mean]edu/[promo_promo]edu = [inc_mean]exp/[promo_promo]exp
  (2)  [inc_mean]edu/[promo_promo]edu = [inc_mean]male/[promo_promo]male

               chi2(2) =        0.61
           Prob > chi2 =        0.7385
```

- Above, we accounted for within-sibling correlation by estimating the cluster–robust VCE.
- We might instead want to fit a multilevel model with sibling-level random intercepts.
- With `gsem`, we can fit a multilevel linear model and a multilevel probit model simultaneously and then test for proportionality of coefficients as we did after using `suest`.

```
. gsem (inc <- edu exp male RI1[famid]) ///
>      (promo <- edu exp male RI2[famid], probit), ///
>          cov(RI1[famid]*RI2[famid]@0) nolog noheader
 ( 1)  [promo]RI2[famid] = 1
 ( 2)  [inc]RI1[famid] = 1
```

	Coefficient	Std. err.	z	P>|z|	[95% conf.	interval]
inc						
edu	2.197144	.2390085	9.19	0.000	1.728696	2.665592
exp	1.470407	.2286072	6.43	0.000	1.022345	1.918468
male	.4213169	.4998765	0.84	0.399	-.558423	1.401057
RI1[famid]	1	(constrained)				
_cons	1.314056	.3226528	4.07	0.000	.6816685	1.946444
promo						
edu	.4773074	.0988786	4.83	0.000	.2835088	.671106
exp	.3829109	.0934125	4.10	0.000	.1998258	.5659959
male	.2164412	.1753531	1.23	0.217	-.1272446	.560127
RI2[famid]	1	(constrained)				
_cons	-.4874952	.1231235	-3.96	0.000	-.7288129	-.2461776

```
--------------------------------------------------------------------------------
var(RI1[famid])|   1.423511   1.243476                       .256929    7.886943
var(RI2[famid])|   .1144566    .196605                      .0039492    3.317199
---------------+----------------------------------------------------------------
     var(e.inc)|   14.54941   1.650875                      11.64829    18.17307
--------------------------------------------------------------------------------

. testnl [inc]edu/[promo]edu = ///
>        [inc]exp/[promo]exp = ///
>        [inc]male/[promo]male

  (1)  [inc]edu/[promo]edu = [inc]exp/[promo]exp
  (2)  [inc]edu/[promo]edu = [inc]male/[promo]male

               chi2(2) =        0.96
           Prob > chi2 =        0.6203
```

Example 7: Multiple-equation finite mixture model

- Beyond standard latent class models, we can use the `lclass()` option to fit finite mixture models where regression models vary across unobserved groups.
- We can apply this to standard path models, such as the mediation model for perceived managerial support, job satisfaction, and job performance that we fit previously.

- Recall that we expect `support` to affect `perform` both directly and indirectly via `satis`.
- We fit this model by typing

```
.   gsem (perform <- satis support) ///
         (satis <- support)
```

- We now allow for this relationship to behave differently across two classes.

```
. gsem (perform <- satis support) (satis <- support), lclass(C 2) nolog noheader
```

	Coefficient	Std. err.	z	P>|z|	[95% conf.	interval]
1.C	(base outcome)					
2.C						
_cons	-.5489265	.3014186	-1.82	0.069	-1.139696	.0418431

Class: 1

	Coefficient	Std. err.	z	P>|z|	[95% conf.	interval]
perform						
satis	.7979916	.0931212	8.57	0.000	.6154775	.9805057
support	.680607	.0520413	13.08	0.000	.578608	.782606
_cons	4.86922	.0774976	62.83	0.000	4.717328	5.021113
satis						
support	.2358343	.0379935	6.21	0.000	.1613683	.3103002
_cons	-.2465063	.0514856	-4.79	0.000	-.3474162	-.1455963
var(e.perform)	.3175168	.0230734			.2753667	.3661187
var(e.satis)	.2346723	.0202454			.1981654	.2779046

Class: 2

	Coefficient	Std. err.	z	P>|z|	[95% conf.	interval]
perform						
satis	.6143178	.0947571	6.48	0.000	.4285972	.8000383
support	.6222295	.0734334	8.47	0.000	.4783026	.7661564
_cons	5.267066	.1032811	51.00	0.000	5.064639	5.469494
satis						
support	.2399658	.0549204	4.37	0.000	.1323238	.3476078
_cons	.4792496	.070222	6.82	0.000	.341617	.6168823
var(e.perform)	.3175168	.0230734			.2753667	.3661187
var(e.satis)	.2346723	.0202454			.1981654	.2779046

- As before, we can use `estat lcprob` to estimate the proportion in each class.

```
. estat lcprob
Latent class marginal probabilities                    Number of obs = 1,500
```

	Margin	Delta-method std. err.	[95% conf.	interval]
C				
1	.6338865	.0699515	.4895408	.7576238
2	.3661135	.0699515	.2423762	.5104592

- We can also use `estat lcmean` to estimate the means of each outcome in each class.

```
. estat lcmean
```

Latent class marginal means   Number of obs = 1,500

	Margin	Delta-method std. err.	z	P>|z|	[95% conf.	interval]
1						
perform	4.8919	.0791127	61.83	0.000	4.736842	5.046958
satis	-.2445095	.0514483	-4.75	0.000	-.3453463	-.1436727
2						
perform	5.285358	.101727	51.96	0.000	5.085977	5.484739
satis	.4812813	.0702846	6.85	0.000	.3435262	.6190365

- Finally, we can use `nlcom` to estimate the indirect effect of `support` on `perform` via `satis` in each class by multiplying coefficients.

```
. nlcom (IE1: _b[perform:1.C#c.satis]*_b[satis:1.C#c.support]) ///
>       (IE2: _b[perform:2.C#c.satis]*_b[satis:2.C#c.support])

        IE1: _b[perform:1.C#c.satis]*_b[satis:1.C#c.support]
        IE2: _b[perform:2.C#c.satis]*_b[satis:2.C#c.support]
```

	Coefficient	Std. err.	z	P>|z|	[95% conf.	interval]
IE1	.1881938	.0359431	5.24	0.000	.1177466	.2586409
IE2	.1474153	.0404982	3.64	0.000	.0680402	.2267903

Example 8: Mixture of different models

- In the previous example, we fit the same model in each latent class. However, we may know that different models are appropriate for different individuals, but we don't know who falls in each group.
- We can use the `lclass()` option in this situation along with numbers prefixing the models to indicate which equation belongs with which class.
- We demonstrate using an artificial dataset on development of calcification after cataract surgery.

```
. webuse lenses, clear
(Simulated calcification data)

. describe

Contains data from https://www.stata-press.com/data/r19/lenses.dta
 Observations:           770                  Simulated calcification data
    Variables:             5                  16 Feb 2025 13:26
--------------------------------------------------------------------------------
Variable      Storage   Display    Value
    name         type    format    label      Variable label
--------------------------------------------------------------------------------
inclength       float   %9.0g                 Incision length (mm)
sex             byte    %9.0g      sexlbl     male=1, female=0
age10           float   %9.0g                 Age (years) divided by 10
t               float   %9.0g                 Time to calcification or to exit
                                                study
fail            byte    %9.0g                 failed=1, didn´t fail=0
--------------------------------------------------------------------------------
Sorted by:
```

- Some individuals experience calcification after surgery.
- Of those who do not have calcification, some are right censored and will eventually have calcification. Others are cured.
- We are intersted in the probability of being cured.
- We fit a survival model for the class that fails and a pointmass distribution for the class that does not fail.
- We first `stset` our data.

```
. stset t, failure(fail)
Survival-time data settings
         Failure event: fail!=0 & fail<.
Observed time interval: (0, t]
     Exit on or before: failure
--------------------------------------------------------------------------
        770  total observations
          0  exclusions
--------------------------------------------------------------------------
        770  observations remaining, representing
        415  failures in single-record/single-failure data
 20,133.467  total analysis time at risk and under observation
                                                At risk from t =         0
                                     Earliest observed entry t =         0
                                          Last observed exit t =        36
```

- We speicify the model for class 1 as a pointmass distribution for those who are cured.
- We specify the model for class 2 as a Weibull model for time to calcification.

```
. gsem (1: fail <-, family(pointmass 0)) ///
>      (2: _t <-  inclength i.sex age10, ///
>            family(weibull, fail(_d) ltruncated(_t0) )), ///
>         lclass(Class 2, base(1)) nolog noheader
```

	Coefficient	Std. err.	z	P>|z|	[95% conf. interval]	
1.Class	(base outcome)					
2.Class						
_cons	1.01863	.2703434	3.77	0.000	.4887664	1.548493

Class: 2

	Coefficient	Std. err.	z	P>|z|	[95% conf. interval]	
_t						
inclength	-.5922698	.2273662	-2.60	0.009	-1.037899	-.1466402
sex						
male	.3314051	.1259957	2.63	0.009	.0844581	.5783522
age10	.1600672	.032798	4.88	0.000	.0957843	.2243502
_cons	-4.939691	.940024	-5.25	0.000	-6.782104	-3.097278
/_t						
ln_p	.4683771	.058332			.3540485	.5827056

- `estat lcprob` reports the probability of being in each class.

```
. estat lcprob
Latent class marginal probabilities                      Number of obs = 770
```

	Margin	Delta-method std. err.	[95% conf.	interval]
Class				
1	.2652944	.0526935	.175304	.3801842
2	.7347056	.0526935	.6198158	.824696

- The probability of being in the cured group is 0.27.

Conclusion

- Generalized structural equation models are very flexible, allowing for modeling many types of outcomes, fitting multilevel models, and including categorical latent variables.
- This allows many extentions of standard structural equation models such as path models, confirmatory factor models, and full structural equation models.
- This also provides some surprising benefits such as comparing parameters across groups or across equations, allowing path models to vary across unobserved groups, and fitting entirely different models for different unobserved classes.
- There are many ways to combine features of `gsem` to fit other interesting models that we have not demonstrated here.