Estimating and interpreting structural equation models in Stata 12

David M. Drukker

Director of Econometrics
Stata

2011 Italian Stata Users Group meeting, Venice
November 17–18, 2011

Purpose and outline

Purpose

To excite structural-equation-model (SEM) devotees by describing part of the new `sem` command and convince traditional simultaneous-equation-model types that the `sem` command is worth investigating

Outline

1. The language of SEM
2. Parameter estimation
   - SUR with observed exogenous variables
   - Recursive (triangular) system with correlated errors
   - SUR with observed exogenous variables and a latent variable
   - Nonrecursive system with a latent variable
3. Postestimation

Variables and Paths

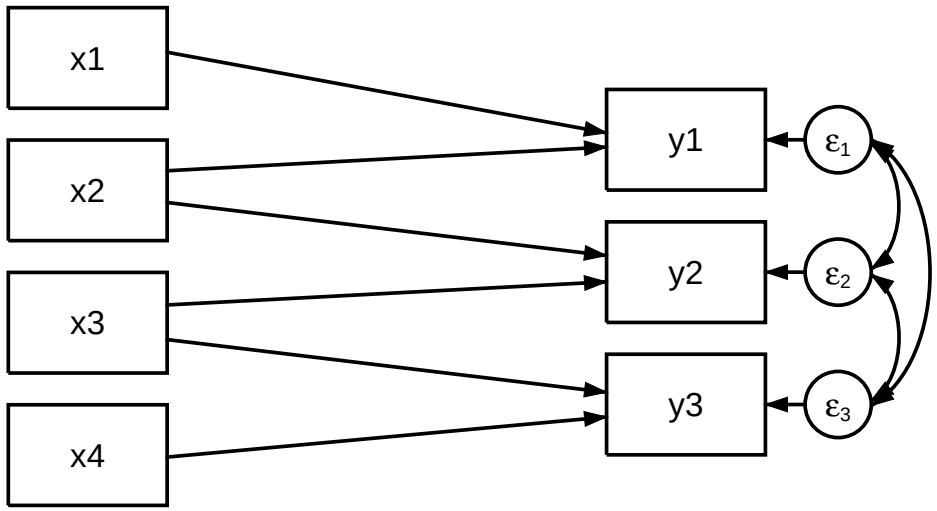
There are five types of variables in SEMs

- A variable is either observed or latent
  - Observed variables are in your dataset
  - Unobserved variables are not in your dataset, but you wish they were
- A variable is either exogenous or endogenous
  - A variable is exogenous if it is determined outside the system
  - A variable is endogenous if is not exogenous
- The concepts give rise to four possibilities
  - Observed exogenous variable, latent exogenous variable, observed endogenous variable, and latent endogenous variable
- Errors are a special type of latent exogenous variables
  - Errors are the random shocks or effects that drive the system
  - Errors are the random effects that cause the outcomes of "observationally equivalent individuals" to differ

Path diagram

- A path diagram is graphical specification of model
- A path diagram is composed of
  - Variables in square or rectangular boxes are observed variables
  - Variables in circles or ellipses are latent variables
  - Straight arrows
    - Each straight arrow indicates that the variable at the base affects the variable at the head
    - When two variables have two arrows that point to each other there is feedback; each one affects the other
  - Curved two-headed arrows indicate that two variables are correlated
  - A number along an arrow represents a constraint

Path diagram



- This is a path diagram for a seemingly unrelated regression (SUR) model with observed exogenous variables

Mathematical description of model

- SUR with observed exogenous variables

$$
\begin{aligned}
y_1 &= \beta_{10} + \beta_{11} x_1 + \beta_{12} x_2 + \epsilon_1 \\
y_2 &= \beta_{20} + \beta_{22} x_2 + \beta_{23} x_3 + \epsilon_2 \\
y_3 &= \beta_{30} + \beta_{33} x_2 + \beta_{34} x_4 + \epsilon_3
\end{aligned}
$$

  where $\boldsymbol{\epsilon} = (\epsilon_1, \epsilon_2, \epsilon_3)'$, $\mathbf{E}[\boldsymbol{\epsilon}] = (0, 0, 0)'$, and $\mathbf{Var}[\boldsymbol{\epsilon}] = \boldsymbol{\Sigma}$

- `sem (y1 <- x1 x2) (y2 <- x2 x3) (y3 <- x3 x4) , cov(e.y2*e.y1 e.y3*e.y2 e.y3*e.y1)`
  alternatively
  `sem (y1 <- x1 x2) (y2 <- x2 x3) (y3 <- x3 x4) , covstructure(e._Endogenous, unstructured)`

Estimate SUR by sem

```
. sem   (y1 <- x1 x2) (y2 <- x2 x3) (y3 <- x3 x4) ,       ///
>         covariance(e.y2*e.y1 e.y3*e.y2 e.y3*e.y1) nolog
Endogenous variables
Observed:  y1 y2 y3
Exogenous variables
Observed:  x1 x2 x3 x4
Structural equation model                       Number of obs      =       500
Estimation method  = ml
Log likelihood     = -6783.5255
```

	OIM Coef.	Std. Err.	z	P>|z|	[95% Conf.	Interval]
Structural						
y1 <-						
x1	.9856651	.0349005	28.24	0.000	.9172614	1.054069
x2	.5498082	.0411897	13.35	0.000	.4690778	.6305385
_cons	.9780043	.0827437	11.82	0.000	.8158297	1.140179
y2 <-						
x2	.3666458	.0443247	8.27	0.000	.2797711	.4535206
x3	1.088846	.0402088	27.08	0.000	1.010038	1.167654
_cons	-1.002962	.0843895	-11.88	0.000	-1.168363	-.837562
y3 <-						
x3	.3069075	.0408562	7.51	0.000	.2268308	.3869841
x4	.7640136	.0396892	19.25	0.000	.6862241	.841803
_cons	1.044546	.0874646	11.94	0.000	.8731183	1.215973
Variance						
e.y1	3.408108	.2158503			3.010253	3.858545
e.y2	3.545391	.2244101			3.131744	4.013674
e.y3	3.823403	.242093			3.377172	4.328596
Covariance						
e.y1						
e.y2	1.949872	.1785632	10.92	0.000	1.599895	2.29985
e.y3	2.151246	.1884359	11.42	0.000	1.781918	2.520573
e.y2						
e.y3	1.940438	.1866187	10.40	0.000	1.574672	2.306204

```
LR test of model vs. saturated: chi2(6)    =      7.40, Prob > chi2 = 0.2855
. estimates store sur_sem
```

Estimate SUR by sureg

```
. sureg (y1 = x1 x2) (y2 = x2 x3) (y3 = x3 x4) , isure nolog  tol(1e-15)
Seemingly unrelated regression, iterated
```

Equation	Obs	Parms	RMSE	"R-sq"	chi2	P
y1	500	2	1.846106	0.6512	1447.37	0.0000
y2	500	2	1.882921	0.6335	1169.28	0.0000
y3	500	2	1.955352	0.4582	644.10	0.0000

	Coef.	Std. Err.	z	P>|z|	[95% Conf.	Interval]
y1						
x1	.9856651	.0348271	28.30	0.000	.9174052	1.053925
x2	.5498082	.0411671	13.36	0.000	.4691222	.6304941
_cons	.9780043	.0827435	11.82	0.000	.81583	1.140179
y2						
x2	.3666458	.0442686	8.28	0.000	.279881	.4534107
x3	1.088846	.0401428	27.12	0.000	1.010167	1.167524
_cons	-1.002962	.084389	-11.88	0.000	-1.168362	-.8375629
y3						
x3	.3069075	.0407619	7.53	0.000	.2270156	.3867993
x4	.7640136	.0395484	19.32	0.000	.6865001	.841527
_cons	1.044546	.0874645	11.94	0.000	.8731185	1.215973

```
. estimates store sur_sureg
```

Results are the same

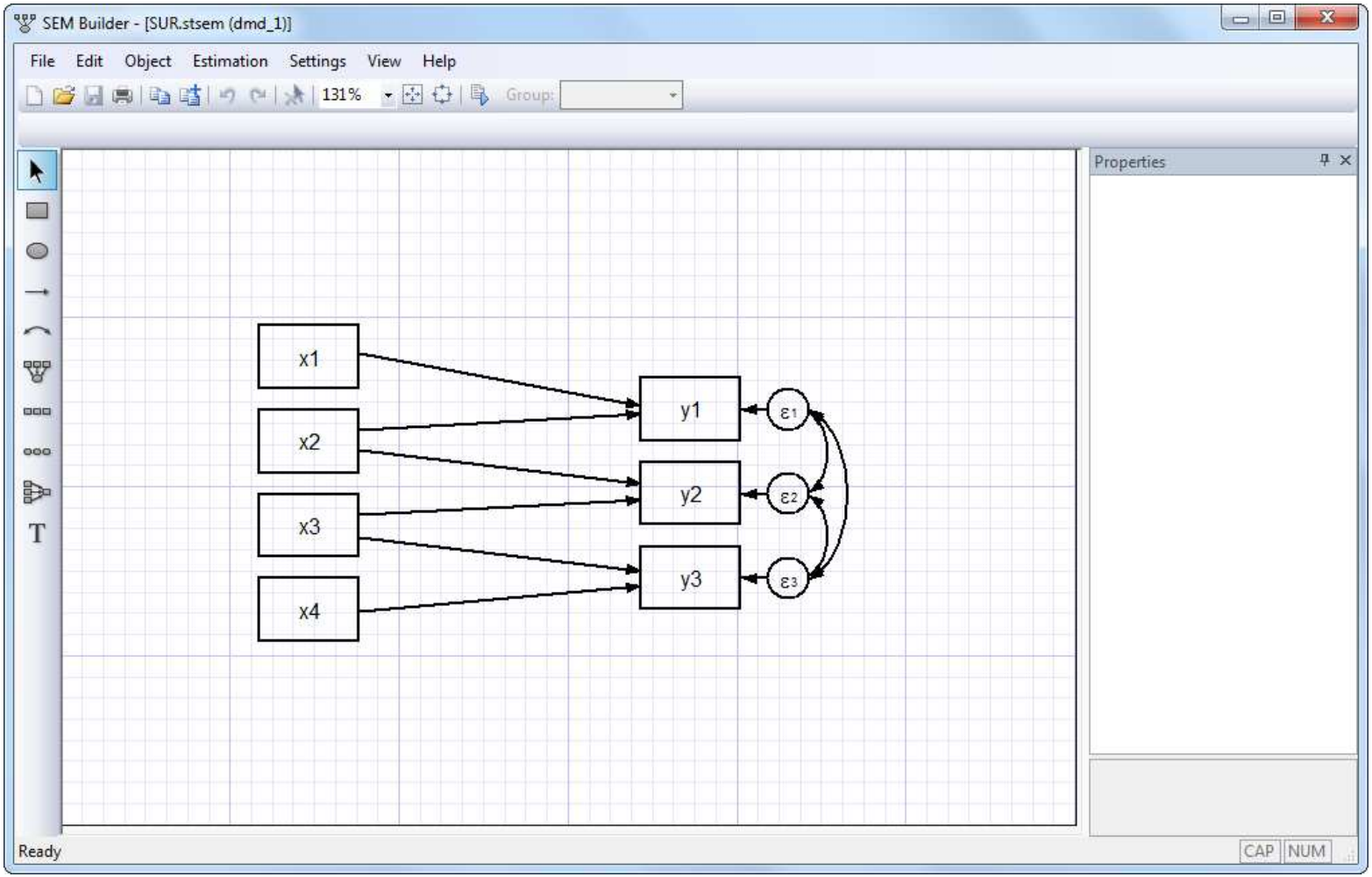
```
. estimates table sur_sem sur_sureg, b se(%7.6g) keep(y1: y2: y3:)
```

Variable	sur_sem	sur_sureg
y1		
x1	.98566508	.98566508
	.0349	.03483
x2	.54980818	.54980818
	.04119	.04117
_cons	.97800427	.97800427
	.08274	.08274
y2		
x2	.36664584	.36664584
	.04432	.04427
x3	1.0888457	1.0888457
	.04021	.04014
_cons	-1.0029623	-1.0029623
	.08439	.08439
y3		
x3	.30690746	.30690746
	.04086	.04076
x4	.76401355	.76401355
	.03969	.03955
_cons	1.0445458	1.0445458
	.08746	.08746

legend: b/se

Sembuilder

- There is an awesome GUI for **sem**

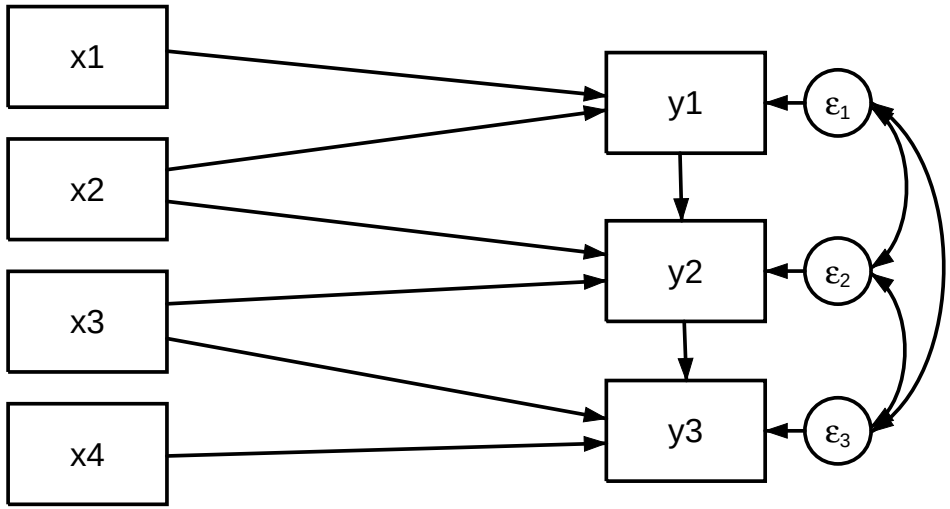


Covariates, errors, and distributions

In all the examples that I discuss

- The analysis is conditional on the exogenous variables
- We assume that the vector of errors, denoted by ϵ, is independently and identically distributed over the observations
- We do not need to assume that the ϵ is normally, or even symmetrically distributed
- Both the Maximum Likehood (ML) and the asymptotically distribution free (ADF) estimators are consistent and asymptotically normally distributed
  - Specify `vce(robust)` with the ML estimator, if the ϵ are not assumed to be normally distributed
  - If the ϵ are normally distributed, the ML estimator is more efficient than the ADF estimator
  - The ADF estimator is a generalized method of moments (GMM) estimator

Path diagram



- Recursive system with correlated errors (SEM language)
  - Sometimes called partially recursive system with correlated errors (SEM language)
- Triangular system with correlated errors (Econometric language)
- The system of equations has a recursive structure, but the errors are correlated so the equation-by-equation ordinary least-squares (OLS) estimator is not consistent.

Mathematical description of model

- Recursive (triangular) system with correlated errors

$$
\begin{aligned}
y_1 &= \beta_{10} + \beta_{11} x_1 + \beta_{12} x_2 + \epsilon_1 \\
y_2 &= \beta_{20} + \gamma_{21} y_1 + \beta_{22} x_2 + \beta_{23} x_3 + \epsilon_2 \\
y_3 &= \beta_{30} + \gamma_{32} y_2 + \beta_{33} x_2 + \beta_{34} x_4 + \epsilon_3
\end{aligned}
$$

  where $\boldsymbol{\epsilon} = (\epsilon_1, \epsilon_2, \epsilon_3)'$, $\mathbf{E}[\boldsymbol{\epsilon}] = (0, 0, 0)'$, and $\mathbf{Var}[\boldsymbol{\epsilon}] = \boldsymbol{\Sigma}$

- `sem (y1 <- x1 x2) (y2 <- y1 x2 x3) (y3 <- y2 x3 x4) , cov(e.y2*e.y1 e.y3*e.y2 e.y3*e.y1)`

Estimate recursive model by sem

```
. sem   (y1 <- x1 x2) (y2 <- y1 x2 x3) (y3 <- y2 x3 x4) ,         ///
>         covariance(e.y2*e.y1 e.y3*e.y2 e.y3*e.y1) nolog
Endogenous variables
Observed:  y1 y2 y3
Exogenous variables
Observed:  x1 x2 x3 x4
Structural equation model                       Number of obs      =       500
Estimation method  = ml
Log likelihood     =  -6882.468
```

	OIM Coef.	Std. Err.	z	P>|z|	[95% Conf.	Interval]
Structural						
y1 <-						
x1	.992947	.0388633	25.55	0.000	.9167763	1.069118
x2	.5402264	.0417589	12.94	0.000	.4583805	.6220723
_cons	.8546342	.0775166	11.03	0.000	.7027045	1.006564
y2 <-						
y1	.5160286	.0463833	11.13	0.000	.425119	.6069381
x2	.5097059	.0627235	8.13	0.000	.38677	.6326417
x3	1.009926	.0429949	23.49	0.000	.9256576	1.094194
_cons	-1.027349	.0983129	-10.45	0.000	-1.220039	-.8346598
y3 <-						
y2	.5732566	.0454244	12.62	0.000	.4842263	.6622868
x3	.2917948	.0729249	4.00	0.000	.1488646	.434725
x4	.8197978	.0444761	18.43	0.000	.7326262	.9069694
_cons	.8690175	.0896196	9.70	0.000	.6933663	1.044669
Variance						
e.y1	2.988624	.1890178			2.640197	3.383033
e.y2	3.886285	.2900963			3.357343	4.498561
e.y3	3.563744	.3279421			2.97562	4.26811
Covariance						
e.y1						
e.y2	1.669049	.2188169	7.63	0.000	1.240175	2.097922
e.y3	1.592503	.2179365	7.31	0.000	1.165355	2.019651
e.y2						
e.y3	1.805499	.3037502	5.94	0.000	1.21016	2.400839

```
LR test of model vs. saturated: chi2(4)    =      0.04, Prob > chi2 = 0.9998
. estimates store sur_sem
```

Estimate recursive model by GLS

- There is a long history in statistics and econometrics of "tricking" readily available estimators to handle more complicated problems
- Using a generalized least squares (GLS) estimator of a triangular SUR model to estimate the parameters of triangular models goes back to [Lahiri and Schmidt(1978)]
- [Prucha(1987)] showed that the standard errors produced by the GLS estimator of a triangular SUR model are not consistent

Estimate recursive model by sureg

```
. sureg (y1 = x1 x2) (y2 = y1 x2 x3) (y3 = y2 x3 x4) , isure nolog  tol(1e-15)
Seemingly unrelated regression, iterated
```

Equation	Obs	Parms	RMSE	"R-sq"	chi2	P
y1	500	2	1.728764	0.7246	1530.52	0.0000
y2	500	3	1.971366	0.8247	2387.49	0.0000
y3	500	3	1.887788	0.8561	2919.81	0.0000

	Coef.	Std. Err.	z	P>|z|	[95% Conf.	Interval]
y1						
x1	.992947	.0374362	26.52	0.000	.9195735	1.066321
x2	.5402264	.0405265	13.33	0.000	.460796	.6196568
_cons	.8546342	.0775078	11.03	0.000	.7027217	1.006547
y2						
y1	.5160286	.0317023	16.28	0.000	.4538932	.5781639
x2	.5097058	.05344	9.54	0.000	.4049654	.6144462
x3	1.009926	.0420154	24.04	0.000	.9275772	1.092275
_cons	-1.02735	.0932221	-11.02	0.000	-1.210061	-.8446377
y3						
y2	.5732566	.0240356	23.85	0.000	.5261477	.6203655
x3	.2917947	.0509012	5.73	0.000	.1920302	.3915593
x4	.8197978	.0419108	19.56	0.000	.7376541	.9019415
_cons	.8690175	.086074	10.10	0.000	.7003156	1.037719

```
. estimates store sur_sureg
```

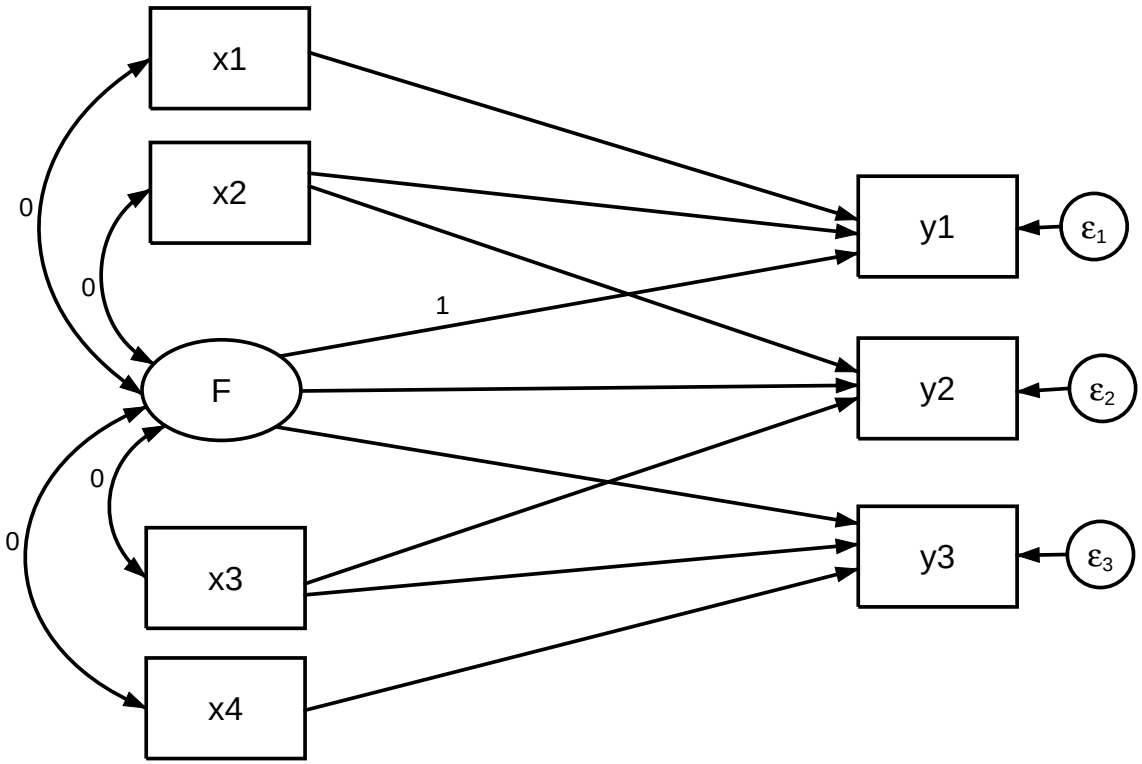
Comparing the results

```
. estimates table sur_sem sur_sureg, b se(%7.6g) keep(y1: y2: y3:)
```

Variable	sur_sem	sur_sureg
y1		
x1	.99294698	.99294699
	.03886	.03744
x2	.54022642	.54022642
	.04176	.04053
_cons	.85463424	.85463424
	.07752	.07751
y2		
y1	.51602855	.51602858
	.04638	.0317
x2	.50970586	.50970583
	.06272	.05344
x3	1.009926	1.009926
	.04299	.04202
_cons	-1.0273495	-1.0273495
	.09831	.09322
y3		
y2	.57325657	.57325658
	.04542	.02404
x3	.29179476	.29179474
	.07292	.0509
x4	.81979779	.81979779
	.04448	.04191
_cons	.8690175	.86901751
	.08962	.08607

legend: b/se

Path diagram



- SUR with observed exogenous variables and a latent variable

Mathematical description of model

- SUR model with observed exogenous variables and a latent variable

$$
\begin{aligned}
y_1 &= \beta_{10} + F + \beta_{11} x_1 + \beta_{12} x_2 + \epsilon_1 \\
y_2 &= \beta_{20} + \rho_2 F + \beta_{22} x_2 + \beta_{23} x_3 + \epsilon_2 \\
y_3 &= \beta_{30} + \rho_3 F + \beta_{33} x_2 + \beta_{34} x_4 + \epsilon_3
\end{aligned}
$$

where $\boldsymbol{\epsilon} = (\epsilon_1, \epsilon_2, \epsilon_3)'$, $\mathbf{E}[\boldsymbol{\epsilon}] = (0, 0, 0)'$, and

$\mathbf{Var}[\boldsymbol{\epsilon}] = \begin{pmatrix} \sigma_1^2 & 0 & 0 \\ 0 & \sigma_2^2 & 0 \\ 0 & 0 & \sigma_3^2 \end{pmatrix}$, $\mathbf{E}[F] = 0$, and $\mathbf{Var}[F] = \sigma_F^2$.

- `sem (y1 <- F x1 x2) (y2 <- F x2 x3) (y3 <- F x3 x4) , cov(F*(x1 x2 x3 x4)@0 )`

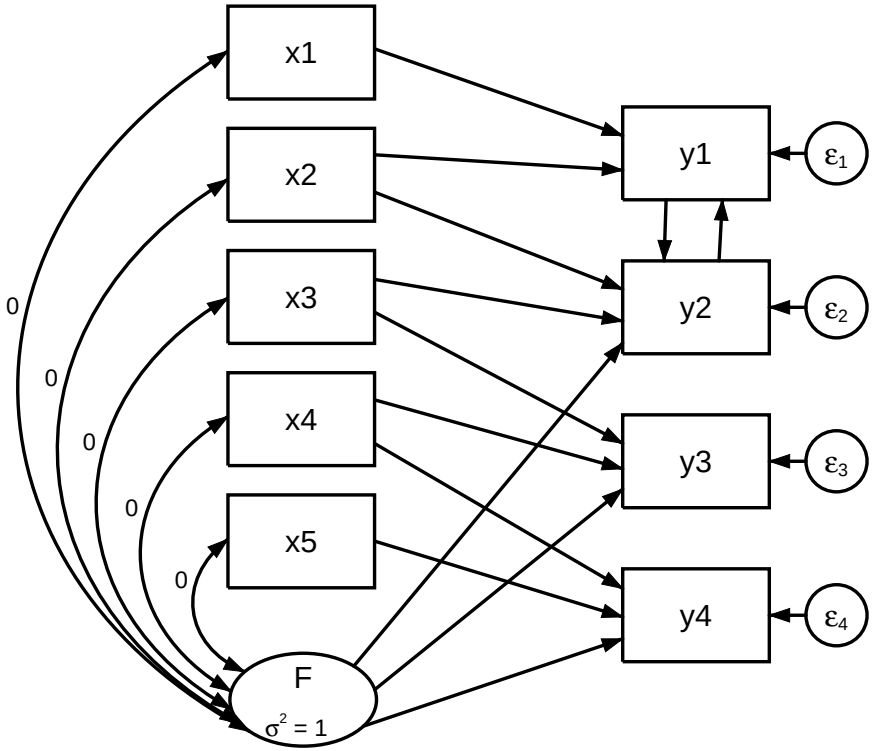
```
. sem   (y1 <- F x1 x2) (y2 <- F x2 x3) (y3 <- F x3 x4) ,         ///
>         covariance(F*(x1 x2 x3 x4)@0 ) nolog
Endogenous variables
Observed:  y1 y2 y3
Exogenous variables
Observed:  x1 x2 x3 x4
Latent:    F
Structural equation model                       Number of obs      =      1000
Estimation method  = ml
Log likelihood     = -13388.953
 ( 1)  [y1]F = 1
 ( 2)  [cov(x1,F)]_cons = 0
 ( 3)  [cov(x2,F)]_cons = 0
 ( 4)  [cov(x3,F)]_cons = 0
 ( 5)  [cov(x4,F)]_cons = 0
```

	OIM Coef.	Std. Err.	z	P>|z|	[95% Conf.	Interval]
Structural						
y1 <-						
x1	.9833928	.0308434	31.88	0.000	.9229408	1.043845
x2	.4752493	.0325475	14.60	0.000	.4114573	.5390413
F	1	(constrained)				
_cons	.9836244	.0564412	17.43	0.000	.8730018	1.094247
y2 <-						
x2	.4352693	.0306857	14.18	0.000	.3751264	.4954122
x3	1.02039	.0278183	36.68	0.000	.9658674	1.074913
F	.8578341	.1012175	8.48	0.000	.6594514	1.056217
_cons	-1.010054	.053389	-18.92	0.000	-1.114695	-.905414
y3 <-						
x3	.2991783	.0281943	10.61	0.000	.2439186	.3544381
x4	.7973739	.030267	26.34	0.000	.7380516	.8566962
F	.6020631	.0701647	8.58	0.000	.4645429	.7395833
_cons	1.086239	.0521731	20.82	0.000	.9839821	1.188497
Variance						
e.y1	1.692668	.1851758			1.366002	2.097452
e.y2	1.751188	.1469865			1.485549	2.064327
e.y3	2.180155	.1151949			1.965674	2.418038
F	1.48224	.2073715			1.126762	1.949868
Covariance						
x1						
F	0	(constrained)				
x2						
F	0	(constrained)				

Variance				
e.y1	1.692668	.1851758	1.366002	2.097452
e.y2	1.751188	.1469865	1.485549	2.064327
e.y3	2.180155	.1151949	1.965674	2.418038
F	1.48224	.2073715	1.126762	1.949868
Covariance				
x1				
F	0	(constrained)		
x2				
F	0	(constrained)		
x3				
F	0	(constrained)		
x4				
F	0	(constrained)		

```
LR test of model vs. saturated: chi2(6)    =      7.07, Prob > chi2 = 0.3144
. estimates store sur_sem
```

Path diagram



- Nonrecursive system with a latent variable

Mathematical description of model

- Simultaneous equation model with observed exogenous variables and a latent variable

$$
\begin{aligned}
y_1 &= \beta_{10} + \gamma_{12} y_2 + \beta_{11} x_1 + \beta_{12} x_2 + \epsilon_1 \\
y_2 &= \beta_{20} + \gamma_{21} y_1 + \rho_2 F + \beta_{22} x_2 + \beta_{23} x_3 + \epsilon_2 \\
y_3 &= \beta_{30} + \rho_3 F + \beta_{33} x_3 + \beta_{34} x_4 + \epsilon_3 \\
y_4 &= \beta_{40} + \rho_4 F + \beta_{44} x_4 + \beta_{45} x_5 + \epsilon_4
\end{aligned}
$$

where $\boldsymbol{\epsilon} = (\epsilon_1, \epsilon_2, \epsilon_3, \epsilon_4)'$, $\mathbf{E}[\boldsymbol{\epsilon}] = (0, 0, 0, 0)'$,

$$
\mathbf{Var}[\boldsymbol{\epsilon}] = \begin{pmatrix} \sigma_1^2 & 0 & 0 & 0 \\ 0 & \sigma_2^2 & 0 & 0 \\ 0 & 0 & \sigma_3^2 & \\ 0 & 0 & 0 & \sigma_4^2 \end{pmatrix} \quad \mathbf{E}[F] = 0, \text{ and } \mathbf{Var}[F] = 1.
$$

- `sem (y1 <- y2 x1 x2) (y2 <- y1 F x2 x3) (y3 <- F x3 x4) (y4 <- F x4 x5), cov(F*(x1 x2 x3 x4 x5)@0 F@1)`

```
. sem   (y1 <- y2 x1 x2) (y2 <- y1 F x2 x3) (y3 <- F x3 x4)       ///
>         (y4 <- F x4 x5), covariance(F*(x1 x2 x3 x4 x5)@0 F@1) nolog
Endogenous variables
Observed:  y1 y2 y3 y4
Exogenous variables
Observed:  x1 x2 x3 x4 x5
Latent:    F
Structural equation model                       Number of obs      =      1000
Estimation method  = ml
Log likelihood     =  -18510.86
 ( 1)  [cov(x1,F)]_cons = 0
 ( 2)  [cov(x2,F)]_cons = 0
 ( 3)  [cov(x3,F)]_cons = 0
 ( 4)  [cov(x4,F)]_cons = 0
 ( 5)  [cov(x5,F)]_cons = 0
 ( 6)  [var(F)]_cons = 1
```

	OIM Coef.	Std. Err.	z	P>|z|	[95% Conf.	Interval]
Structural						
y1 <-						
y2	.7028299	.0067294	104.44	0.000	.6896405	.7160193
x1	.9690673	.0392658	24.68	0.000	.8921077	1.046027
x2	.4794083	.0413284	11.60	0.000	.3984061	.5604105
_cons	1.017156	.048112	21.14	0.000	.9228579	1.111453
y2 <-						
y1	.9965355	.0077294	128.93	0.000	.9813861	1.011685
x2	.5415487	.0429737	12.60	0.000	.4573218	.6257755
x3	.9930621	.0399742	24.84	0.000	.9147141	1.07141
F	1.104189	.0886921	12.45	0.000	.9303558	1.278023
_cons	-.9047846	.0573987	-15.76	0.000	-1.017284	-.7922852
y3 <-						
x3	.2819725	.0294298	9.58	0.000	.2242911	.3396538
x4	.8366874	.0298967	27.99	0.000	.7780909	.8952839
F	.7132992	.0667455	10.69	0.000	.5824804	.8441181
_cons	1.112365	.0520723	21.36	0.000	1.010306	1.214425
y4 <-						
x4	1.00644	.0343146	29.33	0.000	.9391846	1.073695
x5	.7611961	.0331493	22.96	0.000	.6962246	.8261677
F	1.233119	.0947678	13.01	0.000	1.047378	1.418861
_cons	1.131975	.0614426	18.42	0.000	1.01155	1.252401
Variance						
e.y1	2.29826	.1454277			2.030193	2.601722
e.y2	1.961869	.1893448			1.623747	2.370399
e.y3	2.195999	.1171373			1.978008	2.438015
e.y4	2.240033	.2153604			1.855319	2.704521

Variance				
e.y1	2.298255	.1422471	2.035703	2.59467
e.y2	1.961872	.1884104	1.625266	2.368191
e.y3	2.195999	.1171369	1.978009	2.438014
e.y4	2.240033	.2153591	1.855321	2.704517
F	1	(constrained)		
Covariance				
x1				
F	0	(constrained)		
x2				
F	0	(constrained)		
x3				
F	0	(constrained)		
x4				
F	0	(constrained)		
x5				
F	0	(constrained)		

LR test of model vs. saturated: chi2(13) = 12.47, Prob > chi2 = 0.4899

Standard postestimation

- Most standard postestimation features in Stata work after `sem`
  - `test`, `lrtest`, `lincom`, `testnl`, `nlcom`, `predict`, and the `estimates` commands are some important postestimation commands that work after `sem`
- `margins` does not work after `sem` because of the latent variables

Special postestimation

- Some of the important postestimation commands written or modified specifically for `sem`
  - `estat gof`, `estat mindicies`, `estat scoretests`, `estat stdize`, `estat stable`, and `estat teffects`

Direct and indirect effects

- `estat teffects` computes direct effect, indirect effects, total effects and their standard errors
- The direct effect of a variable x on an endogenous variable y is the coefficient on x in the equation for y
  - What is the change in y attributable to a unit change in x, conditional on all other variables in the equation
  - This effect ignores any simultaneous effects
- The total effect of a variable x is the change in an endogenous variable y attributable to a unit change in x after accounting for all the simultaneity in the system
  - Solve the system for the reduced form
  - The total effects are the coefficients in the reduced form specification
- The indirect effect of a variable is the total effect minus the direct effect

Direct effects example

```
. estat teffects, noindirect nototal
Direct effects
```

	OIM Coef.	Std. Err.	z	P>|z|	[95% Conf.	Interval]
Structural						
y1 <-						
y1	0	(no path)				
y2	.7028299	.0067294	104.44	0.000	.6896405	.7160193
x1	.9690673	.0392658	24.68	0.000	.8921077	1.046027
x2	.4794083	.0413284	11.60	0.000	.3984061	.5604105
x3	0	(no path)				
F	0	(no path)				
y2 <-						
y1	.9965355	.0077294	128.93	0.000	.9813861	1.011685
y2	0	(no path)				
x1	0	(no path)				
x2	.5415487	.0429737	12.60	0.000	.4573218	.6257755
x3	.9930621	.0399742	24.84	0.000	.9147141	1.07141
F	1.104189	.0886921	12.45	0.000	.9303558	1.278023
y3 <-						
x3	.2819725	.0294298	9.58	0.000	.2242911	.3396538
x4	.8366874	.0298967	27.99	0.000	.7780909	.8952839
F	.7132992	.0667455	10.69	0.000	.5824804	.8441181
y4 <-						
x4	1.00644	.0343146	29.33	0.000	.9391846	1.073695
x5	.7611961	.0331493	22.96	0.000	.6962246	.8261677
F	1.233119	.0947678	13.01	0.000	1.047378	1.418861

Total effects example

```
. estat teffects, noindirect nodirect
Total effects
```

		OIM Coef.	Std. Err.	z	P>|z|	[95% Conf.	Interval]
Structural							
y1 <-							
	y1	2.337727	.0181321	128.93	0.000	2.302189	2.373266
	y2	2.345855	.0224609	104.44	0.000	2.301832	2.389877
	x1	3.234482	.1155808	27.98	0.000	3.007948	3.461017
	x2	2.870529	.1209948	23.72	0.000	2.633383	3.107674
	x3	2.329579	.1107797	21.03	0.000	2.112455	2.546704
	F	2.590267	.2187261	11.84	0.000	2.161572	3.018963
y2 <-							
	y1	3.326164	.0257986	128.93	0.000	3.275599	3.376728
	y2	2.337727	.0223831	104.44	0.000	2.293857	2.381598
	x1	3.223276	.1316009	24.49	0.000	2.965344	3.481209
	x2	3.402132	.1416452	24.02	0.000	3.124513	3.679752
	x3	3.314571	.1357119	24.42	0.000	3.04858	3.580561
	F	3.685483	.2992404	12.32	0.000	3.098982	4.271983
y3 <-							
	x3	.2819725	.0294298	9.58	0.000	.2242911	.3396538
	x4	.8366874	.0298967	27.99	0.000	.7780909	.8952839
	F	.7132992	.0667455	10.69	0.000	.5824804	.8441181
y4 <-							
	x4	1.00644	.0343146	29.33	0.000	.9391846	1.073695
	x5	.7611961	.0331493	22.96	0.000	.6962246	.8261677
	F	1.233119	.0947678	13.01	0.000	1.047378	1.418861

Random-effects with an endogenous variable

- This example shows how to estimate the parameters of a random-effects model with an endogenous variable
- Doing the estimation with `sem` instead of with `xtivreg` allows the use of `estat teffects` to estimate the total effects
- [Bollen and Brand(2010)] and [Wiggins(2011)] discuss some of these ideas in greater depth

Panel-data long to wide

-
- The trick to estimating panel-data models with `sem` to transform the data to wide format
- In a balanced panel-data analysis, we model

$$\mathbf{y}_i = \mathbf{x}_i\boldsymbol{\beta} + \boldsymbol{\iota} u_i + \boldsymbol{\epsilon}_i$$

  where $\mathbf{y}_i$, $\boldsymbol{\iota}$, and $\boldsymbol{\epsilon}_i$ are all $T \times 1$ vectors, $\mathbf{x}_i$ is a $T \times k$ matrix, and $\boldsymbol{\beta}$ is $k \times 1$ vector
- This mathematical formulation leads us to work with the data in long form

Long data

```
 . use reend, clear
. describe
Contains data from reend.dta
  obs:         3,000
 vars:             6                          2 Nov 2011 13:58
 size:        72,000
```

variable name	storage type	display format	value label	variable label
id	float	%9.0g		
t	float	%9.0g		
x	float	%9.0g		
w	float	%9.0g		
z	float	%9.0g		
y	float	%9.0g		

```
Sorted by:  id  t
. list id t y x if id<=3, sepby(id)
```

	id	t	y	x
1.	1	1	13.05405	.4696761
2.	1	2	5.58284	.0149474
3.	1	3	5.883681	.5247133
4.	2	1	5.293131	.0596235
5.	2	2	5.516943	.0848647
6.	2	3	2.788784	.0867824
7.	3	1	.9604596	.2282464
8.	3	2	2.93892	.8880479
9.	3	3	4.147722	.7269677

Wide data

```
. reshape wide y x w z, i(id) j(t)
(note: j = 1 2 3)
Data                               long   ->   wide
-----------------------------------------------------------------------------
Number of obs.                     3000   ->    1000
Number of variables                   6   ->      13
j variable (3 values)                 t   ->   (dropped)
xij variables:
                                      y   ->   y1 y2 y3
                                      x   ->   x1 x2 x3
                                      w   ->   w1 w2 w3
                                      z   ->   z1 z2 z3
-----------------------------------------------------------------------------
```

```
. list id y1 y2 y3 x1 x2 x3 in 1/3
```

	id	y1	y2	y3	x1	x2	x3
1.	1	13.05405	5.58284	5.883681	.4696761	.0149474	.5247133
2.	2	5.293131	5.516943	2.788784	.0596235	.0848647	.0867824
3.	3	.9604596	2.93892	4.147722	.2282464	.8880479	.7269677

Random-effects model with endogenous variable

$$y_{i1} = x_{i1}\beta + z_{i1}\delta + u_i + \epsilon_{i1} \qquad z_{i1} = x_{i1}\beta + w_{i1}\delta + \eta_{i1}$$
$$y_{i2} = x_{i2}\beta + z_{i2}\delta + u_i + \epsilon_{i2} \qquad z_{i2} = x_{i2}\beta + w_{i2}\delta + \eta_{i2}$$
$$y_{i3} = x_{i3}\beta + z_{i3}\delta + u_i + \epsilon_{i3} \qquad z_{i3} = x_{i3}\beta + w_{i3}\delta + \eta_{i3}$$

- u_i is the unobserved panel-level random effect which is not related to x, z, ϵ, or η
- $\mathbf{E}[\epsilon_{it}] = 0$ for all t,
- $\mathbf{E}[\eta_{it}] = 0$ for all t,
- $\mathbf{E}[\epsilon_{is}\eta_{it}] = \rho$ for all $s = t$, and
- $\mathbf{E}[\epsilon_{is}\eta_{it}] = 0$ for all $s \neq t$

SEM command

```
sem (y1 <- x1@b1 z1@b2 U@1)                                    ///
    (y2 <- x2@b1 z2@b2 U@1)                                    ///
    (y3 <- x3@b1 z3@b2 U@1)                                    ///
    (z1 <- w1@g1 x1@g2)                                        ///
    (z2 <- w2@g1 x2@g2)                                        ///
    (z3 <- w3@g1 x3@g3)                                        ///
    ,                                                          ///
    cov(e.y1*e.z1@rho  e.y2*e.z2@rho  e.y3*e.z3@rho            ///
       U*(x1 x2 x3 w1 w2 w3)@0)
```

```
. sem (y1 <- x1@b1 z1@b2 U@1)                                   ///
>     (y2 <- x2@b1 z2@b2 U@1)                                   ///
>     (y3 <- x3@b1 z3@b2 U@1)                                   ///
>     (z1 <- w1@g1 x1@g2)                                       ///
>     (z2 <- w2@g1 x2@g2)                                       ///
>     (z3 <- w3@g1 x3@g3)                                       ///
>     ,                                                         ///
>     cov(e.y1*e.z1@rho e.y2*e.z2@rho e.y3*e.z3@rho             ///
>        U*(x1 x2 x3 w1 w2 w3)@0)  nolog nocnsreport
Endogenous variables
Observed:  y1 z1 y2 z2 y3 z3
Exogenous variables
Observed:  x1 x2 x3 w1 w2 w3
Latent:    U
Structural equation model                       Number of obs      =      1000
Estimation method  = ml
Log likelihood     = -23174.719
```

	OIM Coef.	Std. Err.	z	P>|z|	[95% Conf.	Interval]
Structural						
y1 <-						
z1	1.005461	.0202605	49.63	0.000	.9657512	1.045171
x1	1.011652	.0352401	28.71	0.000	.9425824	1.080721
U	1	2.11e-17	4.7e+16	0.000	1	1
_cons	-.0615623	.0845403	-0.73	0.466	-.2272582	.1041337
z1 <-						
x1	1.03014	.0317526	32.44	0.000	.9679063	1.092374
w1	1.497454	.0273873	54.68	0.000	1.443776	1.551132
_cons	-.0185884	.0757555	-0.25	0.806	-.1670665	.1298897
y2 <-						
z2	1.005461	.0202605	49.63	0.000	.9657512	1.045171
x2	1.011652	.0352401	28.71	0.000	.9425824	1.080721
U	1	3.00e-17	3.3e+16	0.000	1	1
_cons	.0822569	.0913175	0.90	0.368	-.0967221	.261236
z2 <-						
x2	1.03014	.0317526	32.44	0.000	.9679063	1.092374
w2	1.497454	.0273873	54.68	0.000	1.443776	1.551132
_cons	.0385895	.0763852	0.51	0.613	-.1111227	.1883016
y3 <-						
z3	1.005461	.0202605	49.63	0.000	.9657512	1.045171
x3	1.011652	.0352401	28.71	0.000	.9425824	1.080721
U	1	(constrained)				
_cons	.0538024	.087918	0.61	0.541	-.1185138	.2261185
z3 <-						
x3	1.029447	.0402637	25.57	0.000	.9505318	1.108363

Conclusion

- SEM devotees know that I have only scratched the surface
- Simultaneous-equation types may be interested in including latent variables in their models
  - The postestimation commands, particularly `teffects`, can estimate partial effect parameters and compute specification tests that are not available from other commands for estimating the parameters of simultaneous equation models
  - Even if you are not interested in SEM, you may be interested in `sem`


Bollen, Kenneth A. and Jennie E. Brand. 2010. "A General Panel Model with Random and Fixed Effects: A Structural Equations Approach," *Social Forces*, 89, 1–34.


Lahiri, Kajal and Peter Schmidt. 1978. "A Note on the Consistency of the GLS Estimator in Triangular Structural Systems," *Econometrica*, 46, 1217–1221.


Prucha, Ingmar R. 1987. "The Variance-Covariance Matrix of the Maximum Likelihood Estimator in Triangular Structural Systems: Consistent Estimation," *Econometrica*, 55, 977–978.


Wiggins, Vince. 2011. "Structural equation modeling for those who think they dont care," Tech. rep., Proceedings of the 211 UK Stata Users Group meeting, http://www.stata.com/meeting/uk11/abstracts/UK11_Wiggins.pdf.