Brief Overview of Structural Equation Modeling Using Stata's SEM

Richard Williams, University of Notre Dame, https://www3.nd.edu/~rwilliam/
Last revised April 6, 2015

I am going to use Stata's `sem` commands in this handout. An older handout shows how to do the same things using LISREL. Alan Acock's *Discovering Structural Equation Modeling Using Stata, Revised Edition* is an excellent source for a beginner using sem.

STRUCTURAL AND MEASUREMENT MODELS. We have focused on structural models. Such models assume that all variables are measured without error. Of course, this assumption is often not reasonable. As we saw earlier in the course,

- Random measurement error in the dependent variable does not bias regression coefficients. However, it does result in larger standard errors.
- Random measurement error in the independent variables results in biased estimates. In the case of a bivariate regression, estimates will be biased toward zero. With more IVs, the bias can be upwards or downwards.
- Systematic error, of course, can produce either an upward or downward bias.

Factor analysis is one way of dealing with measurement error. With factor analysis, a large number of items are reduced to a smaller number of factors, or "latent variables". For example, 7 personality measures might be reduced into a single "locus of control" scale. This scale would be more reliable than any of the individual measures that constructed it.

Factor analysis can be either

- exploratory — the computer determines what the underlying factors are
- confirmatory — the researcher specifies what factor structure she thinks underlies the measures, and then tests whether the data are consistent with her hypotheses.

Stata 12 added the `sem` suite of commands. Programs such as `sem` or LISREL make it possible to combine structural equation modeling and confirmatory factor analysis. (I understand programs like AMOS and M-Plus and the `gllamm` addon routine to Stata can do these sorts of things too but I have never used them. These programs may be easier to use and/or cheaper and/or more powerful, so you may want to check them out if you want to do heavy-duty work in this area. For example, some programs can handle ordinal or binary dependent variables, while, at least as of Stata version 12, `sem` cannot.) Some traits of `sem`:

- There is both a measurement model and a structural model.
    - The measurement model indicates how observed indicators are linked to underlying latent variables. (e.g. X1 and X2 may be indicators of Locus of control; X3 and X4 may be indicators of Socio-economic status).
    - The structural model indicates how the latent variables are linked to each other.

- As various sources discuss (e.g. see the Thomson and Williams piece discussed below) having multiple indicators of concepts can help deal with measurement error and thereby produce unbiased estimates of structural effects.

- `sem` can handle a wide array of problems and models. These include
    - Models with measurement error
    - Nonrecursive models
    - Manova-type problems
    - Multiple group comparisons (e.g. you can have separate models for blacks & whites)
    - Tests of constraints (e.g. two or more coefficients equal each other, a subset of coefficients equals zero, parameters are equal across populations)
    - Confirmatory factor analysis models

I'll give just a few examples, not all of which I will talk about in class. It is hard to show in a handout, but `sem` can let you draw the model and will then generate the code for you. This is often the easiest way to go, but the code it generates is not necessarily the clearest or most concise.

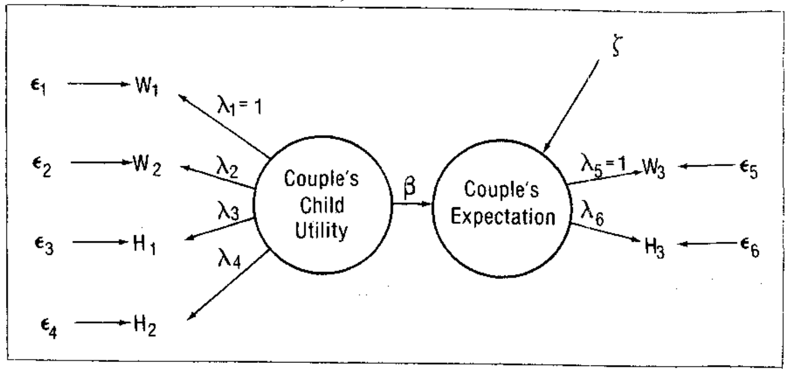
EXAMPLE 1: Measurement and Structural Models Combined. In their classic 1982 paper, "Beyond Wives Family Sociology: A Method for Analyzing Couple Data," Thomson and Williams estimate both measurement and structural parameters in a series of models of couple childbearing expectations. In their data, husbands and their wives were presented with several possible consequences of having another child within 20 months.

- Products of their subjective probability of each consequence (0 = no chance to 10 = certain) and their evaluations of the consequence (-3 = extremely bad thru +3 = extremely good) were constructed to form "subjective expected utilities" of another child. The subjective expected utilities of "a fulfilled family life" (W1 and H1) and "watching another child grow and develop" (W2 and H2) were used as multiple indicators of child utility.

- Also, respondents were asked to estimate the likelihood that the couple would have another child within 20 months (1 = extremely unlikely thru 7 = extremely likely.) Responses of both partners (W3 and H3) were used as multiple indicators of couple childbearing expectations.

Thomson and Williams began by estimating a "couple" model, in which the wife's and husband's responses about the utility of another child are all imperfectly measured indicators of a single latent variable, the couple's child utility. Here is their original diagram for this model:

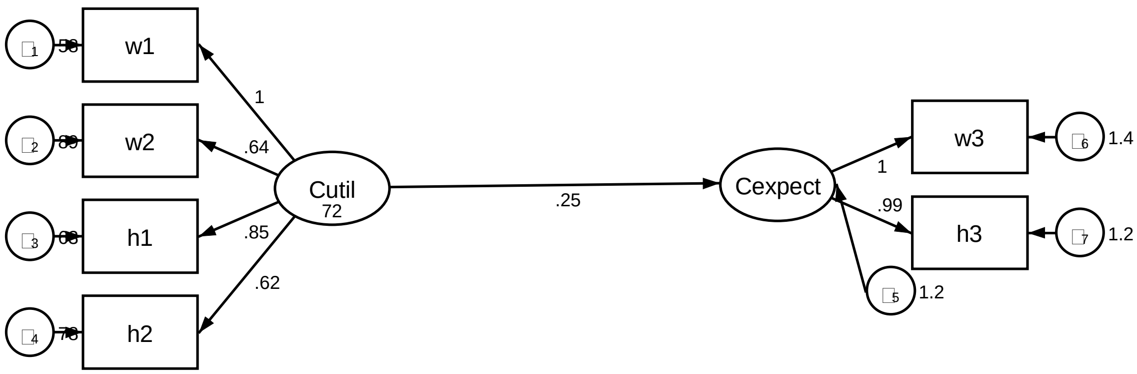
FIGURE 1. COUPLE'S UTILITY OF ANOTHER CHILD AND COUPLE'S CHILDBEARING EXPECTION (VARIABLE LABELS DEFINED IN TABLE 1)



Here is how this model can be estimated with `sem`. The raw data are not available, but the published analyses include the means, standard deviations and correlations for the variables. As in the past, we could use the `corr2data` command to create a pseudo-replication of the data, but the new `ssd` commands (Summary Statistics Data) can now achieve the same purpose. Basically, you first create matrices with the published values and then use the `ssd` commands to tell Stata what the means, correlations and standard deviations are. (I have deleted some of the output that Stata provides along the way.)

```
. * EXAMPLE 1: Measurement and Structural Models Combined
. clear all
. matrix input corr = (1,.47,.46,.312,.628,.596\.47,1,.27,.223,.421,.347\ ///
>         .46,.27,1,.495,.498,.586\.312,.223,.495,1,.381,.422\ ///
>         .628,.421,.498,.381,1,.816\.596,.347,.586,.422,.816,1)
. matrix input means = (11.36,22.34,9.75,18.5,3.64,3.66)
. matrix input sds = (11.45,10.89,10.73,10.30,2.66,2.60)
. ssd init w1 w2 h1 h2 w3 h3
. ssd set observations 340
. * Means were in the paper but not used in the models, so not used here
. *ssd set means (stata) means
. ssd set sd (stata) sds
. ssd set correlations (stata) corr
. ssd list

  Observations = 340

  Means undefined; assumed to be 0

  Standard deviations:
     w1      w2      h1      h2      w3      h3
  11.45   10.89   10.73    10.3    2.66     2.6

  Correlations:
    w1     w2     h1     h2     w3     h3
     1
   .47      1
   .46    .27      1
  .312   .223   .495      1
  .628   .421   .498   .381      1
  .596   .347   .586   .422   .816      1
```

Using Stata's `sem` builder (on the menus, click `Statistics > Structural equation modeling (SEM) > Model building and estimation`, I drew this diagram. Stata filled in the estimates after I told it to run the model. The code that was then generated follows.




```
. sem (Cutil -> Cexpect) (Cutil@1 -> w1) (Cutil -> w2) (Cutil -> h1) ///
>         (Cutil -> h2) (Cexpect@1 -> w3) (Cexpect -> h3), latent(Cutil Cexpect )
```

The latent option tells sem that Cutil (Couple's Child Utility) and Cexpect (Couple's expectations) are the two latent variables. The other parts of the command describe the various paths in the model. Cutil affects Cexpect (the β parameter in the original diagram). The indicators of Cutil are w1,w2, h1 and h2. Cutil@1 says the path from Cutil to w1 is fixed at 1; such constraints are necessary in order to set the scale for the latent variable. You can think of this as meaning that Cutil equals what w1 would equal if w1 were measured without error. Similarly, the indicators for Cexpect are w3 and h3, and Cexpect equals what w3 would equal if w3 were measured without error. The output from the command is as follows.

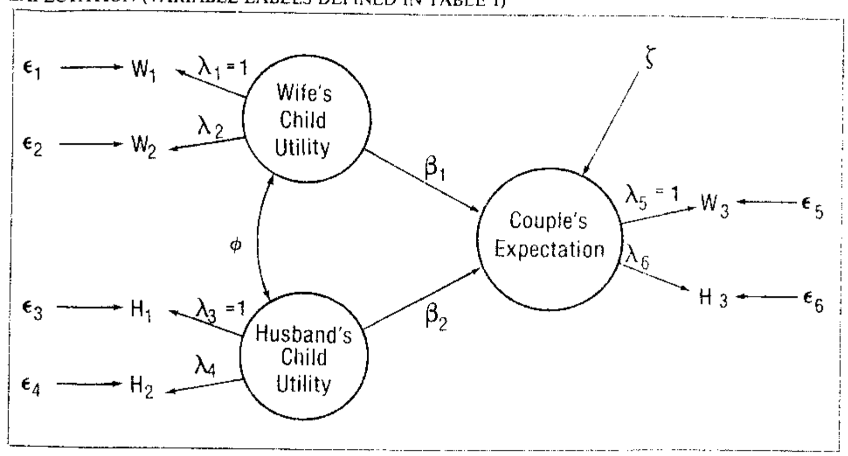
```
Endogenous variables

Measurement:  w1 w2 h1 h2 w3 h3
Latent:       Cexpect

Exogenous variables

Latent:       Cutil

Fitting target model:

Iteration 0:   log likelihood = -6362.6743
Iteration 1:   log likelihood = -6361.2996
Iteration 2:   log likelihood = -6361.2701
Iteration 3:   log likelihood = -6361.2701

Structural equation model                       Number of obs      =       340
Estimation method  = ml
Log likelihood     = -6361.2701
```

```
 ( 1)  [w3]Cexpect = 1
 ( 2)  [w1]Cutil = 1
------------------------------------------------------------------------------
             |                 OIM
             |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
Structural   |
  Cexpect <- |
       Cutil |   .2495012   .0194118    12.85   0.000     .2114548    .2875475
-------------+----------------------------------------------------------------
Measurement  |
  w1 <-      |
       Cutil |          1  (constrained)
  -----------+----------------------------------------------------------------
  w2 <-      |
       Cutil |   .6363228   .0738795     8.61   0.000     .4915216    .7811239
  -----------+----------------------------------------------------------------
  h1 <-      |
       Cutil |   .8486948   .0783366    10.83   0.000      .695158    1.002232
  -----------+----------------------------------------------------------------
  h2 <-      |
       Cutil |   .6240916   .0742715     8.40   0.000      .478522    .7696611
  -----------+----------------------------------------------------------------
  w3 <-      |
     Cexpect |          1  (constrained)
  -----------+----------------------------------------------------------------
  h3 <-      |
     Cexpect |   .9930094   .0482589    20.58   0.000     .8984238    1.087595
-------------+----------------------------------------------------------------
Variance     |
        e.w1 |   58.38456   6.162924                      47.47308      71.804
        e.w2 |   88.95545   7.375962                      75.61241    104.6531
        e.h1 |   62.69452   5.953596                      52.04725    75.51989
        e.h2 |   77.60523   6.478218                      65.89244    91.40003
        e.w3 |    1.38832   .2220985                      1.014652    1.899598
        e.h3 |   1.152595   .2115496                      .8043509    1.651612
   e.Cexpect |    1.16372   .2843919                      .7208259     1.87874
       Cutil |   72.33235   9.919165                      55.28465     94.6369
------------------------------------------------------------------------------
LR test of model vs. saturated: chi2(8)   =     58.91, Prob > chi2 = 0.0000
```

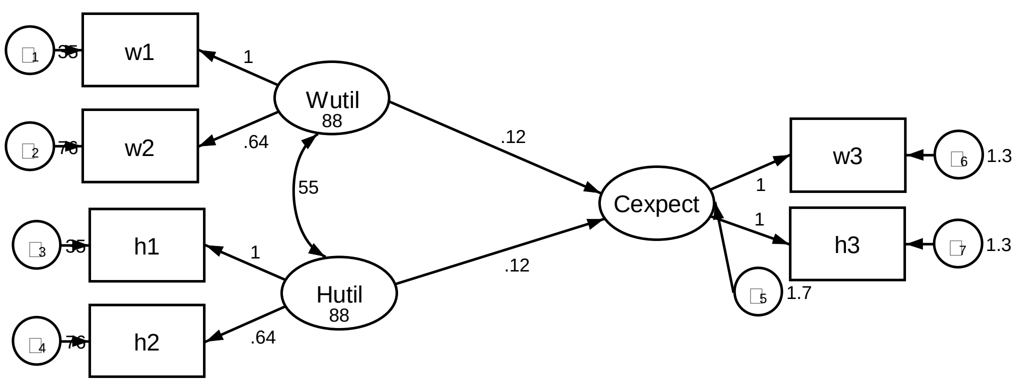
Two things are of particular interest to us. The structural effect of Cutil on Cexpect is .25. We can think of this as the effect that w1 would have on w3 if both were measured without error. The LR test reported at the end tells you how well the model fits the data. The smaller the LR value, the better. [Note that, with 6 observed variables, there are 21 variances and covariances. As the printout shows, only 13 parameters were used in the model, leaving 8 degrees of freedom. Basically, the LR test is testing whether 13 parameters are enough to account for the 21 variances and covariances. The p value says that the fit of the model is not very good, but there are also other ways to assess model fit.]

Thomson and Williams argued that the fit of this model was unacceptable and that rather than having a single couple utility variable, there should be two separate variables, one for husbands and one for wives:

FIGURE 2. WIFE'S AND HUSBAND'S UTILITY OF ANOTHER CHILD AND COUPLE'S CHILDBEARING EXPECTATION (VARIABLE LABELS DEFINED IN TABLE 1)



Also, in their final model (which for some reason they hid in the discussion instead of presenting in the tables) all corresponding parameters between wives and husbands were constrained to be equal. The diagram I created with sem builder and the resulting code it generated is




```
. sem (Wutil@1 -> w1) (Wutil@k2 -> w2) (Wutil@b1 -> Cexpect) (Hutil@1 -> h1) (Hutil@k2
-> h2) (Hutil@b1 -> Cexpect) (Cexpect@1 -> w3) (Cexpect@1 ->
> h3), covstruct(_lexogenous, diagonal) latent(Wutil Hutil Cexpect ) cov( Wutil@v1
Wutil*Hutil e.w1@lx1 e.w2@lx2 Hutil@v1 e.h1@lx1 e.h2@lx2 e.w3@lx3
>  e.h3@lx3) nocapslatent
```

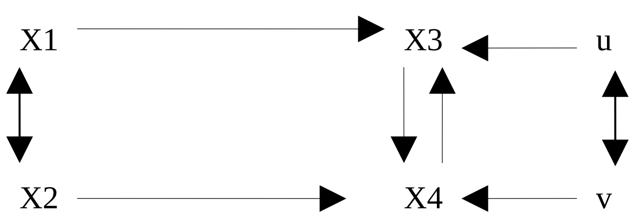
Terms like `(Wutil@k2 -> w2)` and `Hutil@k2 -> h2` mean that all coefficients we have specified as k2 are constrained to be equal. The cov option is specifying the variance/covariance structure. So, Wutil and Hutil can freely covary with each other, and various other variances are unconstrained, but all the other covariances are constrained to be 0.

```
Endogenous variables

Measurement:  w1 w2 h1 h2 w3 h3
Latent:       Cexpect

Exogenous variables

Latent:       Wutil Hutil

Fitting target model:

Structural equation model                         Number of obs      =       340
Estimation method  = ml
Log likelihood     = -6345.5868

 ( 1)  [w3]Cexpect = 1
 ( 2)  [h3]Cexpect = 1
 ( 3)  [w1]Wutil = 1
 ( 4)  [w2]Wutil - [h2]Hutil = 0
 ( 5)  [h1]Hutil = 1
 ( 6)  [Cexpect]Wutil - [Cexpect]Hutil = 0
 ( 7)  [var(e.w1)]_cons - [var(e.h1)]_cons = 0
 ( 8)  [var(e.w2)]_cons - [var(e.h2)]_cons = 0
 ( 9)  [var(e.w3)]_cons - [var(e.h3)]_cons = 0
 (10)  [var(Wutil)]_cons - [var(Hutil)]_cons = 0
------------------------------------------------------------------------------
             |                 OIM
             |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
Structural   |
  Cexpect <- |
       Wutil |   .1175738   .0079036    14.88   0.000      .102083    .1330645
       Hutil |   .1175738   .0079036    14.88   0.000      .102083    .1330645
-------------+----------------------------------------------------------------
Measurement  |
  w1 <-      |
       Wutil |          1  (constrained)
  -----------+----------------------------------------------------------------
  w2 <-      |
       Wutil |   .6447185   .0536132    12.03   0.000     .5396386    .7497984
  -----------+----------------------------------------------------------------
  h1 <-      |
       Hutil |          1  (constrained)
  -----------+----------------------------------------------------------------
  h2 <-      |
       Hutil |   .6447185   .0536132    12.03   0.000     .5396386    .7497984
  -----------+----------------------------------------------------------------
  w3 <-      |
     Cexpect |          1  (constrained)
  -----------+----------------------------------------------------------------
  h3 <-      |
     Cexpect |          1  (constrained)
-------------+----------------------------------------------------------------
```

```
Variance     |
        e.w1 |   35.13733   6.177044                      24.89602    49.59154
        e.w2 |   75.59106   4.771882                      66.79379    85.54699
        e.h1 |   35.13733   6.177044                      24.89602    49.59154
        e.h2 |   75.59106   4.771882                      66.79379    85.54699
        e.w3 |   1.270596   .0974503                       1.09326    1.476698
        e.h3 |   1.270596   .0974503                       1.09326    1.476698
   e.Cexpect |   1.670199   .2734497                      1.211734    2.302128
       Wutil |   87.61817   9.179941                      71.35296    107.5911
       Hutil |   87.61817   9.179941                      71.35296    107.5911
-------------+----------------------------------------------------------------
Covariance   |
  Wutil      |
       Hutil |    55.4943   7.204317     7.70   0.000      41.3741     69.6145
------------------------------------------------------------------------------
LR test of model vs. saturated: chi2(13)  =     27.54, Prob > chi2 = 0.0105
```

This model estimates a total of 8 parameters (remember there are equality constraints on several paraemters), and fits well. Among other things, Thomson and Williams conclude that husbands and wives are not identical in their feelings about the subjective expected utility of children but they are equally influential in determining the couple's expectations for children.

EXAMPLE 2: Nonrecursive Models. The following model has reciprocal effects and is hence nonrecurvice. Using OLS would produce incorrect estimates. Nonrecursive models can be estimated with 2sls or other methods.



We only have single indicators of each X, so no measurement model is used here. This one is pretty easy just to write the code for. First I will estimate using the `reg3` command and 2sls and then `sem`.

```
. reg3 (x3 = x4 x1)(x4 = x3 x2) , 2sls

Two-stage least-squares regression
----------------------------------------------------------------------
Equation          Obs  Parms        RMSE    "R-sq"     F-Stat        P
----------------------------------------------------------------------
x3                500      2    1.779967    0.8009     889.60   0.0000
x4                500      2    4.438984    0.2340     168.62   0.0000
----------------------------------------------------------------------

------------------------------------------------------------------------------
             |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
x3           |
          x4 |  -.2758339   .0238423   -11.57   0.000     -.322621   -.2290468
          x1 |   .4052316   .0096958    41.79   0.000      .386205    .4242582
       _cons |   5.627888     .33629    16.74   0.000     4.967969    6.287808
-------------+----------------------------------------------------------------
x4           |
          x3 |   .6436013   .0651293     9.88   0.000     .5157947     .771408
          x2 |   .4166959   .0229007    18.20   0.000     .3717567    .4616351
       _cons |  -1.859593   1.091455    -1.70   0.089    -4.001414    .2822268
------------------------------------------------------------------------------
Endogenous variables:  x3 x4
Exogenous variables:   x1 x2
------------------------------------------------------------------------------

. use "https://www3.nd.edu/~rwilliam/statafiles/nonrecur.dta", clear
. sem (x1 -> x3) (x2 -> x4) (x3 -> x4) (x4 -> x3), cov( e.x4*e.x3)

Endogenous variables

Observed:  x3 x4

Exogenous variables

Observed:  x1 x2

Fitting target model:

Iteration 0:   log likelihood = -5966.0177
Iteration 1:   log likelihood = -5966.0177

Structural equation model                       Number of obs      =       500
Estimation method  = ml
Log likelihood     = -5966.0177
```

```
------------------------------------------------------------------------------
             |                 OIM
             |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
Structural   |
  x3 <-      |
          x4 |  -.2758339   .0237707   -11.60   0.000    -.3224236   -.2292441
          x1 |   .4052316   .0096667    41.92   0.000     .3862852    .4241779
       _cons |   5.627888   .3352796    16.79   0.000     4.970752    6.285024
  -----------+----------------------------------------------------------------
  x4 <-      |
          x3 |   .6436013   .0649336     9.91   0.000     .5163338    .7708688
          x2 |   .4166959   .0228319    18.25   0.000     .3719463    .4614456
       _cons |  -1.859593   1.088176    -1.71   0.087    -3.992378    .2731915
-------------+----------------------------------------------------------------
Variance     |
        e.x3 |   3.149273   .2030317                      2.775453    3.573443
        e.x4 |   19.58635    1.54716                      16.77705    22.86606
-------------+----------------------------------------------------------------
Covariance   |
  e.x3       |
        e.x4 |  -3.002073   .5543294    -5.42   0.000    -4.088538   -1.915607
------------------------------------------------------------------------------
LR test of model vs. saturated: chi2(0)   =      0.00, Prob > chi2 =      .
```

Also, Duncan-Haller-Portes presented a model of peer influence, where peers had reciprocal influence on each other.

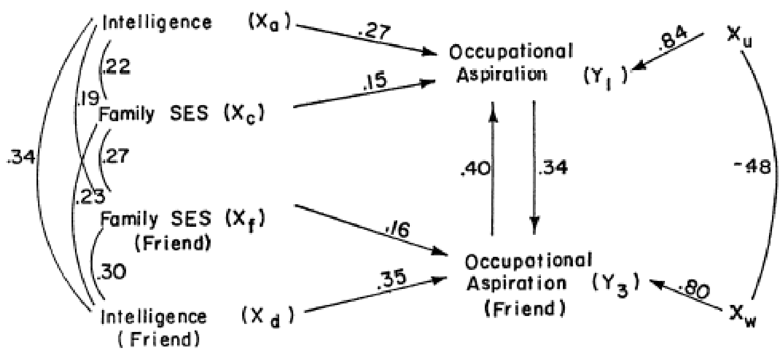


FIG. 2.—Model II

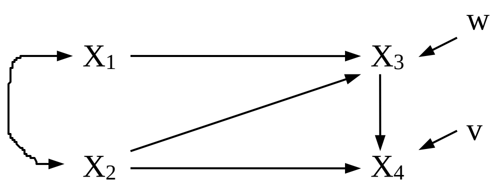
Using the published information in their paper, this model is pretty easy to estimate using `sem`.

```
. * Duncan Haller Portes p. 8
. * A slight variation of this example using same data is in the Stata help
. clear all
. ssd init rintelligence rparasp rses roccasp redasp ///
>        bfintelligence bfparasp bfses bfoccasp bfedasp
. ssd set observations 329
. ssd set corr ///
>  1.0000 \ ///
>   .1839 1.0000 \ ///
>   .2220  .0489 1.0000 \ ///
>   .4105  .2137  .3240 1.0000 \ ///
>   .4043  .2742  .4047  .6247 1.0000 \ ///
>   .3355  .0782  .2302  .2995  .2863 1.0000 \ ///
>   .1021  .1147  .0931  .0760  .0702  .2087 1.0000 \ ///
>   .1861  .0186  .2707  .2930  .2407  .2950 -.0438 1.0000 \ ///
>   .2598  .0839  .2786  .4216  .3275  .5007  .1988  .3607 1.0000 \ ///
```

```
>   .2903  .1124  .3054  .3269  .3669  .5191  .2784  .4105  .6404 1.0000
. sem (rintelligence -> roccasp) (rses -> roccasp) (bfintelligence -> bfoccasp) ///
>         (bfses -> bfoccasp) (roccasp -> bfoccasp) (bfoccasp -> roccasp), ///
>         cov( e.roccasp*e.bfoccasp)

Endogenous variables

Observed:  roccasp bfoccasp

Exogenous variables

Observed:  rintelligence rses bfintelligence bfses

Fitting target model:

Iteration 0:   log likelihood = -2619.6916
Iteration 1:   log likelihood = -2619.1002
Iteration 2:   log likelihood = -2619.0915
Iteration 3:   log likelihood = -2619.0914

Structural equation model                       Number of obs      =       329
Estimation method  = ml
Log likelihood     = -2619.0914

------------------------------------------------------------------------------
                   |                 OIM
                   |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------------+----------------------------------------------------------------
Structural         |
  roccasp <-       |
          bfoccasp |   .4079437    .104743     3.89   0.000     .2026512    .6132362
     rintelligence |    .251426   .0538545     4.67   0.000     .1458732    .3569789
              rses |   .1749922   .0460249     3.80   0.000      .084785    .2651993
  -----------------+----------------------------------------------------------------
  bfoccasp <-      |
           roccasp |    .348331   .1258765     2.77   0.006     .1016175    .5950444
    bfintelligence |   .3276121   .0580873     5.64   0.000      .213763    .4414612
             bfses |   .1862807   .0454284     4.10   0.000     .0972427    .2753187
-------------------+----------------------------------------------------------------
Variance           |
         e.roccasp |    .706912   .0590185                      .6002061    .8325882
        e.bfoccasp |   .6476102   .0543616                      .5493666    .7634227
-------------------+----------------------------------------------------------------
Covariance         |
  e.roccasp        |
        e.bfoccasp |  -.3321255   .1236722    -2.69   0.007    -.5745186   -.0897324
------------------------------------------------------------------------------------
LR test of model vs. saturated: chi2(2)   =      4.08, Prob > chi2 = 0.1297
```

The estimates are very similar to the published results.

Example 3: Decomposing Correlations. We talked at length about how to decompose the correlation between two variables into direct and indirect effects. Consider the following model. Assume that all the X's are standardized, i.e. have mean 0 and variance 1. Also assume that changes in X1 cannot produce changes in X2, and changes in X2 cannot produce changes in X1.



The correlation matrix is

```
. corr
(obs=1000)

             |       x1       x2       x3       x4
-------------+------------------------------------
          x1 |   1.0000
          x2 |   0.6000   1.0000
          x3 |   0.5400   0.5800   1.0000
          x4 |   0.5700   0.7900   0.7900   1.0000
```

Sem can estimate this model and, by using the `estat teffect` command, decompose the correlations into direct and indirect effects.

```
. * EXAMPLE 3: Decomposing Correlations
. clear all
. ssd init x1 x2 x3 x4
. ssd set observations 1000
. ssd set corr (ltd) 1 .60 1 .54 .58 1 .57 .79 .79 1
. sem (x1 x2 -> x3) (x2 x3 -> x4)

Endogenous variables

Observed:  x3 x4

Exogenous variables

Observed:  x1 x2

Fitting target model:

Iteration 0:   log likelihood = -4419.8481
Iteration 1:   log likelihood = -4419.8481

Structural equation model                       Number of obs      =      1000
Estimation method  = ml
Log likelihood     = -4419.8481
```

```
------------------------------------------------------------------------------
             |                 OIM
             |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
Structural   |
  x3 <-      |
          x1 |         .3   .0307713     9.75   0.000     .2396893    .3603107
          x2 |         .4   .0307713    13.00   0.000     .3396893    .4603107
  -----------+----------------------------------------------------------------
  x4 <-      |
          x3 |         .5   .0177892    28.11   0.000     .4651338    .5348662
          x2 |         .5   .0177892    28.11   0.000     .4651338    .5348662
-------------+----------------------------------------------------------------
Variance     |
        e.x3 |    .605394    .027074                        .554589    .6608532
        e.x4 |     .20979   .0093821                       .1921843    .2290085
------------------------------------------------------------------------------
LR test of model vs. saturated: chi2(1)   =      0.00, Prob > chi2 = 1.0000
```

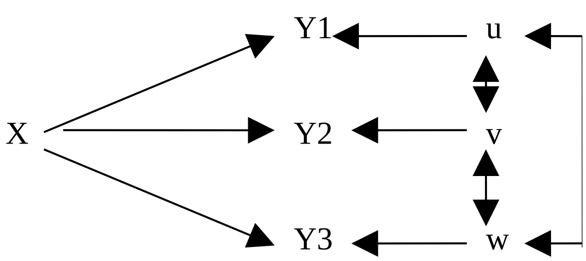
. estat teffects

```
Direct effects
------------------------------------------------------------------------------
             |                 OIM
             |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
Structural   |
  x3 <-      |
          x1 |         .3   .0307713     9.75   0.000     .2396893    .3603107
          x2 |         .4   .0307713    13.00   0.000     .3396893    .4603107
  -----------+----------------------------------------------------------------
  x4 <-      |
          x3 |         .5   .0177892    28.11   0.000     .4651338    .5348662
          x1 |          0  (no path)
          x2 |         .5   .0177892    28.11   0.000     .4651338    .5348662
------------------------------------------------------------------------------


Indirect effects
------------------------------------------------------------------------------
             |                 OIM
             |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
Structural   |
  x3 <-      |
          x1 |          0  (no path)
          x2 |          0  (no path)
  -----------+----------------------------------------------------------------
  x4 <-      |
          x3 |          0  (no path)
          x1 |        .15    .016285     9.21   0.000     .1180821    .1819179
          x2 |         .2   .0169515    11.80   0.000     .1667758    .2332242
------------------------------------------------------------------------------
```

```
Total effects
------------------------------------------------------------------------------
             |                 OIM
             |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
Structural   |
  x3 <-      |
          x1 |         .3   .0307713     9.75   0.000     .2396893    .3603107
          x2 |         .4   .0307713    13.00   0.000     .3396893    .4603107
  -----------+----------------------------------------------------------------
  x4 <-      |
          x3 |         .5   .0177892    28.11   0.000     .4651338    .5348662
          x1 |        .15    .016285     9.21   0.000     .1180821    .1819179
          x2 |         .7   .0213769    32.75   0.000      .658102     .741898
------------------------------------------------------------------------------
```

Hence, `sem` can do some of the decomposition of effects that you have previously done by hand. In complicated models, such decompositions are difficult to compute manually. Knowing the total effect of a variable can be useful, since it tells you how much a 1 unit change in an IV will change the expected value of a DV.

Example 4: Using sem for Manova. Sometimes we are interested in situations where X variables affect multiple dependent variables.



You could estimate such a model using the `manova` and `mvreg` commands:

```
. use https://www3.nd.edu/~rwilliam/statafiles/blwh.dta, clear
. quietly manova income educ jobexp = black
. mvreg

Equation          Obs  Parms        RMSE    "R-sq"          F        P
----------------------------------------------------------------------
income            500      2    7.768778    0.2520   167.7605   0.0000
educ              500      2    3.698475    0.1385     80.066   0.0000
jobexp            500      2    4.931661    0.0526   27.66301   0.0000

------------------------------------------------------------------------------
             |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
income       |
     1.black |     -11.25   .8685758   -12.95   0.000    -12.95652   -9.543475
       _cons |      30.04   .3884389    77.34   0.000     29.27682    30.80318
-------------+----------------------------------------------------------------
```

```
educ         |
     1.black |       -3.7    .413502    -8.95   0.000    -4.512424   -2.887576
       _cons |       13.9   .1849237    75.17   0.000     13.53667    14.26333
-------------+----------------------------------------------------------------
jobexp       |
     1.black |       -2.9   .5513765    -5.26   0.000    -3.983311   -1.816689
       _cons |       14.1   .2465831    57.18   0.000     13.61553    14.58447
------------------------------------------------------------------------------
```

Using `sem` (the `covstructure` option allows the residuals for the three dependent variables to be freely correlated),

```
. sem black -> income educ jobexp, covstructure(e._En, unstructured)

Endogenous variables

Observed:  income educ jobexp

Exogenous variables

Observed:  black

Fitting target model:

Iteration 0:   log likelihood = -4474.1119
Iteration 1:   log likelihood = -4474.1119

Structural equation model                       Number of obs      =       500
Estimation method  = ml
Log likelihood     = -4474.1119

------------------------------------------------------------------------------
             |                 OIM
             |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
Structural   |
  income <-  |
       black |     -11.25   .8668369   -12.98   0.000    -12.94897   -9.551031
       _cons |      30.04   .3876613    77.49   0.000      29.2802     30.7998
  -----------+----------------------------------------------------------------
  educ <-    |
       black |       -3.7   .4126742    -8.97   0.000    -4.508827   -2.891173
       _cons |       13.9   .1845535    75.32   0.000     13.53828    14.26172
  -----------+----------------------------------------------------------------
  jobexp <-  |
       black |       -2.9   .5502727    -5.27   0.000    -3.978515   -1.821485
       _cons |       14.1   .2460894    57.30   0.000     13.61767    14.58233
-------------+----------------------------------------------------------------
```

```
Variance     |
    e.income |    60.1125   3.801848                      53.10435    68.04551
      e.educ |     13.624   .8616574                      12.03566    15.42195
    e.jobexp |     24.224    1.53206                      21.39987    27.42083
-------------+----------------------------------------------------------------
Covariance   |
  e.income   |
      e.educ |    22.2856    1.62211    13.74   0.000     19.10632    25.46488
    e.jobexp |     7.9032   1.742771     4.53   0.000     4.487431    11.31897
  -----------+----------------------------------------------------------------
  e.educ     |
    e.jobexp |      -4.28    .834681    -5.13   0.000    -5.915945   -2.644055
------------------------------------------------------------------------------
LR test of model vs. saturated: chi2(0)   =      0.00, Prob > chi2 =      .
```

Example 5: Using sem for Group Comparisons. We are often interested in making comparisons across groups. For example, we have previously worked with examples like this:

```
. use "https://www3.nd.edu/~rwilliam/statafiles/gender.dta"
. reg income educ jobexp if !female

      Source |       SS       df       MS              Number of obs =     225
-------------+------------------------------           F(  2,   222) =  210.87
       Model |  19350.4582     2  9675.22912           Prob > F      =  0.0000
    Residual |  10185.7638   222  45.8818188           R-squared     =  0.6551
-------------+------------------------------           Adj R-squared =  0.6520
       Total |   29536.222   224  131.858134           Root MSE      =  6.7736

------------------------------------------------------------------------------
      income |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
        educ |   .8195378   .1070818     7.65   0.000     .6085108    1.030565
      jobexp |   1.384972   .0895246    15.47   0.000     1.208545    1.561398
       _cons |  -.9294128    1.49777    -0.62   0.536     -3.88108    2.022254
------------------------------------------------------------------------------

. reg income educ jobexp if female

      Source |       SS       df       MS              Number of obs =     275
-------------+------------------------------           F(  2,   272) =  120.03
       Model |  5276.94296     2  2638.47148           Prob > F      =  0.0000
    Residual |  5979.19312   272  21.9823276           R-squared     =  0.4688
-------------+------------------------------           Adj R-squared =  0.4649
       Total |  11256.1361   274  41.0807886           Root MSE      =  4.6885

------------------------------------------------------------------------------
      income |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
        educ |   1.525582   .1004096    15.19   0.000     1.327903    1.723261
      jobexp |  -.0049199   .0773587    -0.06   0.949    -.1572178    .1473779
       _cons |   5.470545   1.589722     3.44   0.001     2.340821    8.600269
------------------------------------------------------------------------------
```

We've shown various ways to test whether effects differ across groups, and if so how they differ. `sem` offers another alternative. First, we will show how `sem` can replicate the above results (note the use of the `group` option; that causes separate models to be estimated for each group).

```
. *** No constraints across groups
. sem (educ -> income) (jobexp -> income), group(female)

Endogenous variables

Observed:  income

Exogenous variables

Observed:  educ jobexp

Fitting target model:

Iteration 0:   log likelihood = -4327.8267
Iteration 1:   log likelihood = -4327.8267

Structural equation model                       Number of obs      =       500
Grouping variable  = female                     Number of groups   =         2
Estimation method  = ml
Log likelihood     = -4327.8267

------------------------------------------------------------------------------
             |                 OIM
             |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
Structural   |
  income <-  |
    educ     |
        male |   .8195378   .1063656     7.70   0.000     .6110651     1.02801
      female |   1.525582   .0998604    15.28   0.000     1.329859    1.721305
    jobexp   |
        male |   1.384972   .0889258    15.57   0.000      1.21068    1.559263
      female |  -.0049199   .0769356    -0.06   0.949    -.1557109     .145871
    _cons    |
        male |  -.9294128   1.487752    -0.62   0.532    -3.845353    1.986527
      female |   5.470545   1.581027     3.46   0.001     2.371789    8.569301
-------------+----------------------------------------------------------------
Variance     |
  e.income   |
        male |   45.27006   4.268102                      37.63215    54.45818
      female |   21.74252   1.854208                      18.39582    25.69808
------------------------------------------------------------------------------
LR test of model vs. saturated: chi2(0)   =      0.00, Prob > chi2 =      .

. est store m1
```

Note that these are the same as the coefficient estimates we got running separate regressions. We can now estimate a model in which only the intercepts are allowed to differ.

```
. reg income educ jobexp female

      Source |       SS       df       MS              Number of obs =     500
-------------+------------------------------           F(  3,   496) =  189.85
       Model |  24326.2478     3  8108.74928           Prob > F      =  0.0000
    Residual |   21184.389   496  42.7104618           R-squared     =  0.5345
-------------+------------------------------           Adj R-squared =  0.5317
       Total |  45510.6369   499  91.2036811           Root MSE      =  6.5353

------------------------------------------------------------------------------
      income |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
        educ |   1.281368   .0803805    15.94   0.000      1.12344    1.439296
      jobexp |   .7738483   .0652862    11.85   0.000     .6455767      .90212
      female |  -4.071767   .5990074    -6.80   0.000    -5.248671   -2.894862
       _cons |   2.511457   1.269321     1.98   0.048     .0175474    5.005367
------------------------------------------------------------------------------

. *** Pooled model; only intercepts allowed to differ
. sem (educ -> income) (jobexp -> income), group(female) ginvariant(scoef serrvar)
```

The `ginvariant` option on the `sem` command specifies which values are allowed to differ across groups. In this case, the coefficients and the error variance is being constrained to be equal across groups (but not the constant).

```
Endogenous variables

Observed:  income

Exogenous variables

Observed:  educ jobexp

Fitting target model:

Iteration 0:   log likelihood = -4632.8768
Iteration 1:   log likelihood = -4547.3852
Iteration 2:   log likelihood = -4429.6844
Iteration 3:   log likelihood = -4412.6934
Iteration 4:   log likelihood = -4412.1075
Iteration 5:   log likelihood = -4412.1073

Structural equation model                       Number of obs      =       500
Grouping variable  = female                     Number of groups   =         2
Estimation method  = ml
Log likelihood     = -4412.1073
```

```
 ( 1)  [income]0bn.female#c.educ - [income]1.female#c.educ = 0
 ( 2)  [income]0bn.female#c.jobexp - [income]1.female#c.jobexp = 0
 ( 3)  [var(e.income)]0bn.female - [var(e.income)]1.female = 0
------------------------------------------------------------------------------
             |                 OIM
             |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
Structural   |
  income <-  |
    educ     |
         [*] |   1.281368   .0800583    16.01   0.000     1.124456    1.438279
    jobexp   |
         [*] |   .7738483   .0650245    11.90   0.000     .6464026     .901294
    _cons    |
        male |   2.511455   1.264233     1.99   0.047      .033604    4.989306
      female |  -1.560305   1.151915    -1.35   0.176    -3.818017    .6974067
-------------+----------------------------------------------------------------
Variance     |
  e.income   |
         [*] |   42.36871   2.679629                      37.42921    47.96008
------------------------------------------------------------------------------
Note: [*] identifies parameter estimates constrained to be equal across
      groups.
LR test of model vs. saturated: chi2(3)   =    168.56, Prob > chi2 = 0.0000
```

. est store m2

We can also estimate models in which even the intercepts aren't allowed to differ.

. reg income educ jobexp

```
      Source |       SS       df       MS              Number of obs =     500
-------------+------------------------------           F(  2,   497) =  239.86
       Model |  22352.7545     2  11176.3773           Prob > F      =  0.0000
    Residual |  23157.8824   497  46.5953368           R-squared     =  0.4912
-------------+------------------------------           Adj R-squared =  0.4891
       Total |  45510.6369   499  91.2036811           Root MSE      =  6.8261

------------------------------------------------------------------------------
      income |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
        educ |   1.309229   .0838474    15.61   0.000      1.14449    1.473968
      jobexp |   .8533107   .0670888    12.72   0.000     .7214982    .9851233
       _cons |  -1.076636   1.205717    -0.89   0.372    -3.445568    1.292296
------------------------------------------------------------------------------
```

. sem (educ -> income) (jobexp -> income), group(female) ginvariant(scoef serrvar scons)

```
Endogenous variables

Observed:  income

Exogenous variables

Observed:  educ jobexp

Fitting target model:

Structural equation model                       Number of obs      =       500
Grouping variable  = female                     Number of groups   =         2
Estimation method  = ml
Log likelihood     =  -4434.375
```

```
 ( 1)  [income]0bn.female#c.educ - [income]1.female#c.educ = 0
 ( 2)  [income]0bn.female#c.jobexp - [income]1.female#c.jobexp = 0
 ( 3)  [var(e.income)]0bn.female - [var(e.income)]1.female = 0
 ( 4)  [income]0bn.female - [income]1.female = 0
------------------------------------------------------------------------------
             |                 OIM
             |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
Structural   |
  income <-  |
    educ     |
         [*] |   1.309229   .0835954    15.66   0.000     1.145385    1.473073
    jobexp   |
         [*] |   .8533107   .0668872    12.76   0.000     .7222143    .9844072
    _cons    |
         [*] |  -1.076636   1.202095    -0.90   0.370    -3.432699    1.279427
-------------+----------------------------------------------------------------
Variance     |
  e.income   |
         [*] |   46.31576   2.929266                      40.91609    52.42803
------------------------------------------------------------------------------
Note: [*] identifies parameter estimates constrained to be equal across
      groups.
LR test of model vs. saturated: chi2(4)   =    213.10, Prob > chi2 = 0.0000

. est store m3
```

Previously, we did things like F tests to test constraints. Now we can use LR chi-square contrasts. So, contrasting Model 2 (only the constant is allowed to differ across groups) with Model 3 (even the constant is not allowed to differ) we get

```
. lrtest m2 m3

Likelihood-ratio test                                 LR chi2(1)  =     44.54
(Assumption: m3 nested in m2)                         Prob > chi2 =    0.0000
```

We would reject the hypothesis that the constants are the same for the two groups.