# Nonrecursive Models – Highlights

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

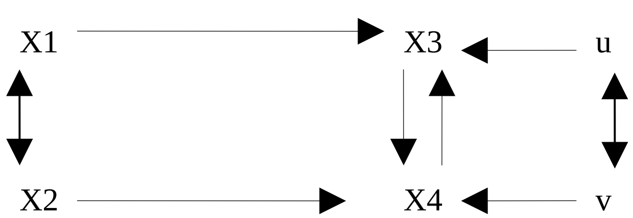
This lecture borrows heavily from Duncan's Introduction to Structural Equation Models and from William D. Berry's Nonrecursive Causal Models. There is a longer version of this handout that goes into much more depth but it is probably overkill for a basic understanding.]

## Introduction

We have previously talked about recursive models. In recursive models, the causal flows all go in one direction, e.g. if X1 affects X2, then X2 does not directly or indirectly also affect X1. Further, we assumed that the disturbance in an equation was uncorrelated with any of the independent variables in the equation. For example, if Y is regressed on X1 and X2, the error term for Y is assumed to be uncorrelated with both X1 and X2. If this assumption is violated and OLS is used to estimate the model, the estimates of the coefficients will be biased.

The assumption that the residual term for Y is uncorrelated with the Xs might be violated if, say, variables were omitted from the model that affected Y that were also correlated with the Xs that were in the model. We previously discussed this as a problem of omitted variable bias, but it can also be thought of as a violation of the OLS requirement that the residual terms must be uncorrelated with the Xs. Berry discusses instances of where such problems might occur. (These are probably the most commonly addressed sorts of problems in the literature today, and eventually I will include some good examples of them.)

Another situation in which assumptions will be violated is when there is reciprocal causation. Consider the following:




In this model, X1 and X2 are *exogenous* variables (their values are determined outside the model) while X3 and X4 are *endogenous* (their values are determined within the model). There are reciprocal effects between X3 and X4. The residuals, u and v, are also correlated.

Note that, in this model, v is correlated with X3, because v affects X4 which in turn affects X3, i.e. v is an indirect cause of X3. Hence, if OLS is used to estimate the regression of X4 on X2 and X3, the assumption that the residual v is uncorrelated with X2 and X3 is violated. Similarly, when X3 is regressed on X1 and X4, OLS assumptions are violated because u is an indirect cause of X4 and hence is correlated with it.  Procedures besides OLS must be used if we want to get correct parameter estimates.

## Estimation of Non-Recursive Models: 2 Stage Least Squares.

There are various ways of estimating this nonrecursive model (e.g. instrumental variables, indirect least squares, LISREL models). For now, I will focus on a technique called 2 stage least squares (2SLS). 2SLS is best done with a single program that handles all the steps. If each step is done separately, the coefficients will be correct but the standard errors will be wrong. To make clear what is going on though, I will show how each step can be estimated separately.

Conceptually, the procedure is as follows:

- Regress each endogenous variable on *all* exogenous variables (in this case, regress X3 on X1 and X2, and regress X4 on X1 and X2). Use the OLS parameter estimates to compute predicted values for X3 and X4:

$$\hat{X}_3 = b^*_{31}X_1 + b^*_{32}X_2$$
$$\hat{X}_4 = b^*_{41}X_1 + b^*_{42}X_2$$

Note that X3-hat and X4-hat will *not* be correlated with the error terms in the model, e.g. since X1 and X2 are not correlated with u and v, and since X3-hat and X4-hat are computed from X1 and X2, X3-hat will *not* be correlated with v and X4-hat will *not* be correlated with u.

In Stata, we could do the first stage as follows:

```
. use https://www3.nd.edu/~rwilliam/statafiles/nonrecur.dta, clear
. quietly reg x3 x1 x2
. predict x3hat if e(sample)
(option xb assumed; fitted values)
. quietly reg x4 x1 x2
. predict x4hat if e(sample)
(option xb assumed; fitted values)
```

- In the second stage of 2SLS, any endogenous variable $X_j$ serving as an explanatory variable in one of the structural equations is replaced by the corresponding predicted variable computed in the first step. In the present case, we estimate the regressions

$$X_3 = \beta_{31}X_1 + \beta_{34}\hat{X}_4 + u$$

$$X_4 = \beta_{42}X_2 + \beta_{43}\hat{X}_3 + v$$

Given these substitutions, each explanatory variable in the modified structural equations can be assumed uncorrelated with the error terms in the model. Hence, you can use OLS to estimate the parameters of the revised structural equations. Using Stata for step 2,

```
. reg x3 x1 x4hat

      Source |       SS       df       MS              Number of obs =     500
-------------+------------------------------           F(  2,   497) =  617.03
       Model |  5636.98124     2  2818.49062           Prob > F      =  0.0000
    Residual |  2270.21876   497  4.56784458           R-squared     =  0.7129
-------------+------------------------------           Adj R-squared =  0.7117
       Total |      7907.2   499  15.8460922           Root MSE      =  2.1373

------------------------------------------------------------------------------
          x3 |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
          x1 |   .4052316    .011642    34.81   0.000      .382358    .4281052
       x4hat |  -.2758339   .0286281    -9.64   0.000    -.3320809   -.2195868
       _cons |   5.627888   .4037919    13.94   0.000     4.834539    6.421238
------------------------------------------------------------------------------

. reg x4 x2 x3hat

      Source |       SS       df       MS              Number of obs =     500
-------------+------------------------------           F(  2,   497) =  268.95
       Model |   6644.9822     2   3322.4911           Prob > F      =  0.0000
    Residual |   6139.8178   497  12.3537581           R-squared     =  0.5198
-------------+------------------------------           Adj R-squared =  0.5178
       Total |     12784.8   499  25.6208417           Root MSE      =  3.5148

------------------------------------------------------------------------------
          x4 |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
          x2 |   .4166959   .0181328    22.98   0.000     .3810696    .4523223
       x3hat |   .6436013   .0515694    12.48   0.000     .5422804    .7449223
       _cons |  -1.859593   .8642149    -2.15   0.032    -3.557558    -.161628
```

2SLS estimators are biased but consistent; that is, as the sample gets larger and larger, the expected values of the 2SLS estimators get closer and closer to the population parameters.

The standard errors of 2SLS estimators are partially a function of the degree to which the variables created in the first stage are similar to the endogenous variables they replace. Ceterus Paribus, the higher the correlation between the predicted variables and the original endogenous variables, the more efficient the parameters produced by 2SLS. The reason we use all (as opposed to some) of the exogenous variables as independent variables in the first stage regressions is because we want to construct variables as similar as possible to the endogenous variables while still making certain that the new variables are uncorrelated with the error terms in the equations.

As described, 2SLS is a procedure involving two separate stages of OLS analysis. Fortunately, Stata and other packages will now do 2SLS as a one step procedure, avoiding the problems of the 2 step OLS approach. Stata has various commands that will do two stage (and also three stage) least squares. These include the `ivregress` and `reg3` commands (see Stata's help for complete details on syntax). `reg3` is a little bit easier to use with models involving reciprocal causation so I will focus on it.

```
. reg3  (x3 = x1 x4) (x4 = x2 x3), 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           |
          x1 |   .4052316   .0096958    41.79   0.000      .386205    .4242582
          x4 |  -.2758339   .0238423   -11.57   0.000     -.322621   -.2290468
       _cons |   5.627888     .33629    16.74   0.000     4.967969    6.287808
-------------+----------------------------------------------------------------
x4           |
          x2 |   .4166959   .0229007    18.20   0.000     .3717567    .4616351
          x3 |   .6436013   .0651293     9.88   0.000     .5157947     .771408
       _cons |  -1.859593   1.091455    -1.70   0.089    -4.001414    .2822268
------------------------------------------------------------------------------
Endogenous variables:  x3 x4
Exogenous variables:   x1 x2
------------------------------------------------------------------------------
```

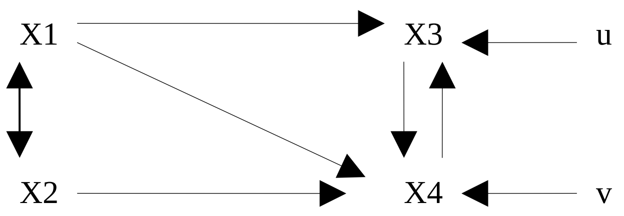
Note that the coefficient estimates are identical to what we got before, but the standard errors are different. This is because, when we do each step separately, the 2nd step estimation does not take into account the fact that some of the variables are regression estimates rather than observed values. The default option of 3 stage least squares produces the same coefficient estimates in this case but slightly different standard errors. 3sls combines two-stage least squares (2SLS) with seemingly unrelated regressions (SUR), i.e. it takes into account the fact that there are multiple equations and that the residuals for those equations may be correlated with each other. 3sls is probably slightly better but, at least in the examples used here, it doesn't seem to matter much.

Incidentally, suppose we just ignored the fact that the residuals were correlated with the Xs and ran an OLS regression of X4 on X1 and X3:

```
. reg x4 x2 x3

      Source |       SS       df       MS              Number of obs =     500
-------------+------------------------------           F(  2,   497) =  151.05
       Model |  4833.29715     2  2416.64858           Prob > F      =  0.0000
    Residual |  7951.50285   497  15.9989997           R-squared     =  0.3781
-------------+------------------------------           Adj R-squared =  0.3755
       Total |     12784.8   499  25.6208417           Root MSE      =  3.9999

------------------------------------------------------------------------------
          x4 |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
          x2 |   .3405887   .0200306    17.00   0.000     .3012336    .3799438
          x3 |   .1275494   .0480987     2.65   0.008     .0330474    .2220513
       _cons |   6.193688   .8318055     7.45   0.000       4.5594    7.827977
------------------------------------------------------------------------------
```

Note that the estimated effect of X3 on X4 is much smaller with OLS (.1275) than it is with 2sls (.644). Hence, in this particular case, failure to take into account that OLS assumptions are violated in this model would lead to a serious underestimate of the effect of X3 on X4. The nature of any biases will vary on a model by model basis though (e.g. if we regress X3 on X1 and X4 the OLS estimates aren't that much different from what you get with 2sls).

The Problem of Underidentification. Unfortunately, estimating nonrecursive models is not just as simple as using a different Stata command. In order to estimate the model it must be identified. Some models, whether true or not, are impossible to estimate. For example, consider this model:




We have now added a path from X1 to X4. Let's see what happens when we try to estimate it.

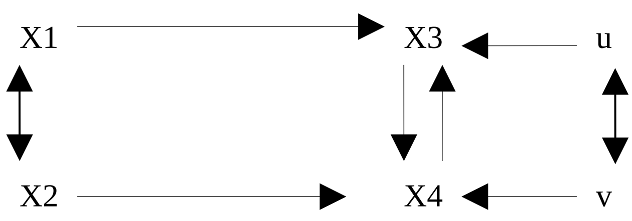
```
. reg3 (x4 = x3 x2 x1) (x3 = x4 x1)
Equation is not identified -- does not meet order conditions
    Equation x4:  x4  x3 x2 x1
    Exogenous variables:   x2 x1
r(481);
```

Why is Stata complaining about the X4 equation?  Let's again try doing the steps separately to gain insight into what is happening.

```
. * Stage 1: Compute x3hat
. quietly reg x3 x1 x2
. predict x3hat if e(sample)
(option xb assumed; fitted values)

. * Second Stage: regress X4 on X1, X2, and X3hat

. reg x4 x1 x2 x3hat
note: x3hat omitted because of collinearity

      Source |       SS       df       MS              Number of obs =     500
-------------+------------------------------           F(  2,   497) =  268.95
       Model |  6644.98219     2  3322.49109           Prob > F      =  0.0000
    Residual |  6139.81781   497  12.3537582           R-squared     =  0.5198
-------------+------------------------------           Adj R-squared =  0.5178
       Total |     12784.8   499  25.6208417           Root MSE      =  3.5148

------------------------------------------------------------------------------
          x4 |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
          x1 |   .2214876    .017747    12.48   0.000     .1866192    .2563559
          x2 |   .3538738   .0166604    21.24   0.000     .3211403    .3866072
       x3hat |          0  (omitted)
       _cons |   1.496801   .6215556     2.41   0.016        .2756    2.718001
------------------------------------------------------------------------------
```

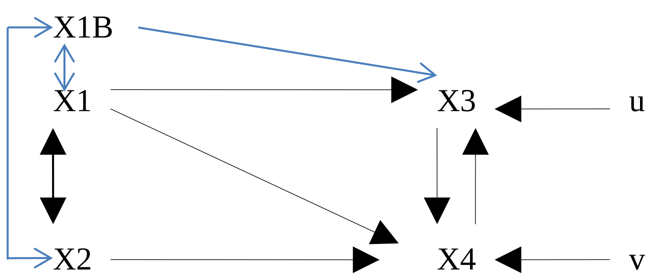
We see that we have a problem of perfect collinearity, i.e. x3hat is perfectly correlated with x1 and x2. Recall that, in the first stage of 2sls, X3 is regressed on X1 and X2 and the predicted value for X3-hat is computed. In the 2$^{nd}$ stage, X4 is regressed on X1, X2, and X3-hat. But herein lies the problem: X3-hat was computed using X1 and X2 (i.e. is a weighted sum of those variables), so when X1, X2 and X3Hat are all in the same model there is a problem of perfect multicollinearity and the model is not identified.

How then do we avoid the problem of underidentification? Suppose Xi and Xj each affect each other (in this case X3 and X4). *For the Xj equation to be identified, there must be at least one predetermined variable that directly affects Xi but not Xj.* This variable is the "instrument" for Xi (or instruments if there is more than one such variable). *Similarly, for the Xi equation to be identified, there must be at least one variable that directly affects Xj but not Xi.* In the present example, X2 affects X4 but not X3, hence the X3 equation is identified. However, every variable that affects X3 also affects X4, hence the X4 equation is not identified. Conversely, in the earlier example,




X2 affected X4 but not X3, and X1 affected X3 but not X4. Hence, as drawn, underidentification is not a problem with this model.

From the above, there would seem to be a straightforward solution to the identification problem. If the Xj equation is underidentified, simply add predetermined variables to the Xi equation but not to the Xj equation. That is, you simply need to add variables in the "right" place. For example, in our underidentified model, it would seem that all we have to do is add a variable X1B that affects X3 but not X4:




However, this is much harder than it sounds.

- *The added variables must have a significant direct effect on X3.* Adding a variable whose expected value is zero is the same as not adding the variable in the first place. Adding weak or extraneous variables may make the model appear to be identified, but in reality they won't solve your problem if their effects are very weak or nonexistent.

  Put another way, *the added variables must make sense theoretically*. If we add a variable to the X3 equation, it should be the case that we think this variable affects X3. If we don't think it has an effect, then its expected value is zero, which means it does us no good to add it.

- Perhaps even more difficult, *we must believe that any added variables have indirect effects on X4, but do not have direct effects on X4.* That is, we have to believe that X3 is the mechanism through which the added variable affects X4, and that once X3 is controlled for, the added variable has no direct effect on X4. It can be quite difficult to think of such variables.

Some examples of where this might make sense:

- Supply and demand — rainfall might affect the supply of agricultural products but not directly affect the demand for them. Per capita income might affect demand but not directly affect supply.

- Peer influence — Peer 1's aspirations may affect Peer 2's aspirations, and vice versa. Peer 1 may be directly influenced by her parent's socio-economic status (SES), but her parent's SES may have no direct effect on her friend's aspiration. Similarly, Peer 2 is directly affected by her parent's SES, but her parent's SES has no direct effect on Peer 1. Ergo, in this case, the respective parents' SES (as well as possibly other background variables of each peer) serve as the instruments.

Here is such an example from **Peer Influences on Aspirations: A Reinterpretation**, Otis Dudley Duncan, Archibald O. Haller, Alejandro Portes, *American Journal of Sociology*, Vol. 74, No. 2. (Sep., 1968), pp. 119-137. Diagram is on p. 126. The study collected data from both respondents and their friends. The model states that peers have reciprocal influence on each other's occupational aspirations. Each peer is directly affected by his own intelligence and family SES, but is only indirectly affected by the intelligence and family SES of his friend.

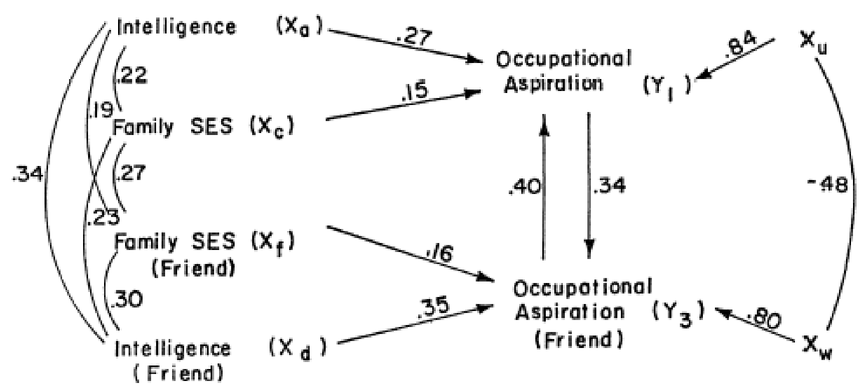



Fig. 2.—Model II

The published means, correlations and standard deviations can be used to reproduce these estimates. We use the `corr2data` command to create a pseudo-replication of the data. We then estimate the model using 2sls (which is apparently what Duncan, Haller and Portes used; if we use 3sls both the coefficients and the standard errors are slightly different).

```
. clear all
. matrix input corr =
(1,.1839,.222,.4105,.4043,.3355,.1021,.1861,.2598,.2903\.1839,1,.0489,.2137,.2742,.078
2,.1147,.0186,.0839,.1124\.222,.0489,1,.
>
324,.4047,.2302,.0931,.2707,.2786,.3054\.4105,.2137,.324,1,.6247,.2995,.076,.293,.4216
,.3269\.4043,.2742,.4047,.6247,1,.2863,.0702,.2407,.3275,.36
>
69\.3355,.0782,.2302,.2995,.2863,1,.2087,.295,.5007,.5191\.1021,.1147,.0931,.076,.0702
,.2087,1,-.0438,.1988,.2784\.1861,.0186,.2707,.293,.2407,.29
> 5,-
.0438,1,.3607,.4105\.2598,.0839,.2786,.4216,.3275,.5007,.1988,.3607,1,.6404\.2903,.112
4,.3054,.3269,.3669,.5191,.2784,.4105,.6404,1)

. corr2data rintelligence rparasp rses roccasp redasp bfintelligence bfparasp bfses bfoccasp bfedasp, n(329) corr(corr)
(obs 329)

. reg3 (roccasp = rintelligence rses bfoccasp) (bfoccasp =  bfses bfintelligence roccasp), 2sls

Two-stage least-squares regression
----------------------------------------------------------------------
Equation          Obs  Parms        RMSE    "R-sq"     F-Stat        P
----------------------------------------------------------------------
roccasp           329      3    .8449421    0.2926      39.53   0.0000
bfoccasp          329      3    .8084131    0.3524      52.76   0.0000
----------------------------------------------------------------------

--------------------------------------------------------------------------------
               |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
---------------+----------------------------------------------------------------
roccasp        |
 rintelligence |   .2721328   .0525467     5.18   0.000     .1689511    .3753145
          rses |   .1512026   .0536377     2.82   0.005     .0458786    .2565266
      bfoccasp |   .4033882   .1043116     3.87   0.000     .1985599    .6082165
         _cons |   5.09e-09   .0465832     0.00   1.000    -.0914716    .0914717
---------------+----------------------------------------------------------------
bfoccasp       |
         bfses |   .1566602   .0544491     2.88   0.004     .0497428    .2635776
bfintelligence |   .3520896   .0550489     6.40   0.000     .2439944    .4601848
       roccasp |   .3418886   .1247791     2.74   0.006     .0968699    .5869073
         _cons |  -3.33e-09   .0445693    -0.00   1.000    -.0875171    .0875171
--------------------------------------------------------------------------------
Endogenous variables:  roccasp bfoccasp
Exogenous variables:   rintelligence rses bfses bfintelligence
------------------------------------------------------------------------------
```

## Appendix (Optional):
## Estimation of Non-Recursive Models with Structural Equation Modeling (sem)

The last two models can also be easily estimated using the `sem` command.

Example 1. First, for our simple 4 variable nonrecursive model,

```
. 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 =      .
```

Example 2. Using the published information in their paper, the Duncan-Haller-Portes model of peer influence, where peers had reciprocal influence on each other, is pretty easy to estimate using `sem`. In the code `ssd` stands for Summary Statistics Data; when used with `sem`, it is an alternative to creating a pseudo-replication with `corr2data`. The cov option tells Stata that the residuals for the two dependent variables are freely correlated.

```
. * 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

Summary statistics data initialized.  Next use, in any order,

    ssd set observations (required)
        It is best to do this first.

    ssd set means (optional)
        Default setting is 0.

    ssd set variances or ssd set sd (optional)
        Use this only if you have set or will set correlations and, even then, this is
optional but highly recommended.  Default setting is 1.

    ssd set covariances or ssd set correlations (required)

. ssd set observations 329
  (value set)

    Status:
                       observations:    set
                              means:  unset
                    variances or sd:  unset
        covariances or correlations:  unset (required to be set)

. 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
  (values set)

    Status:
                       observations:    set
                              means:  unset
                    variances or sd:  unset
        covariances or correlations:    set

. sem (bfintelligence bfses roccasp -> bfoccasp) ///
>         (rintelligence rses bfoccasp -> roccasp), ///
>         cov( e.roccasp*e.bfoccasp)

Endogenous variables

Observed:  roccasp bfoccasp

Exogenous variables

Observed:  bfintelligence bfses rintelligence rses

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, with the differences being due to the fact that a different estimation method (maximum likelihood) was used. The chi-square test at the end suggests that no important paths have been omitted.