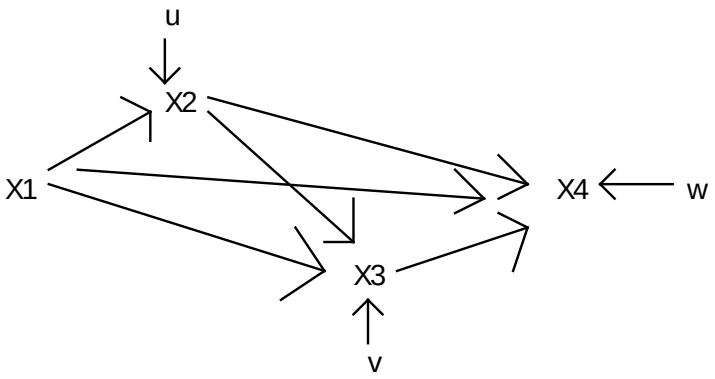
Nonrecursive models (Extended Version)

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

NOTE: This lecture borrows heavily from Duncan's Introduction to Structural Equation Models and from William D. Berry's Nonrecursive Causal Models. There is a shorter version of this handout that leaves out a lot of details but may be easier to follow.

Advantages and Disadvantages of Recursive Models. We have previously considered recursive models such as the following:



Recursive models meet the following conditions:

- Models are hierarchical. All causal effects in the model are "unidirectional" in nature, i.e. no two variables in the model are reciprocally related, either directly or indirectly. Hence, the first endogenous variable is affected only by the exogenous variables. The 2nd endogenous variable is affected only by the exogenous variables and the first endogenous variable; and so on.

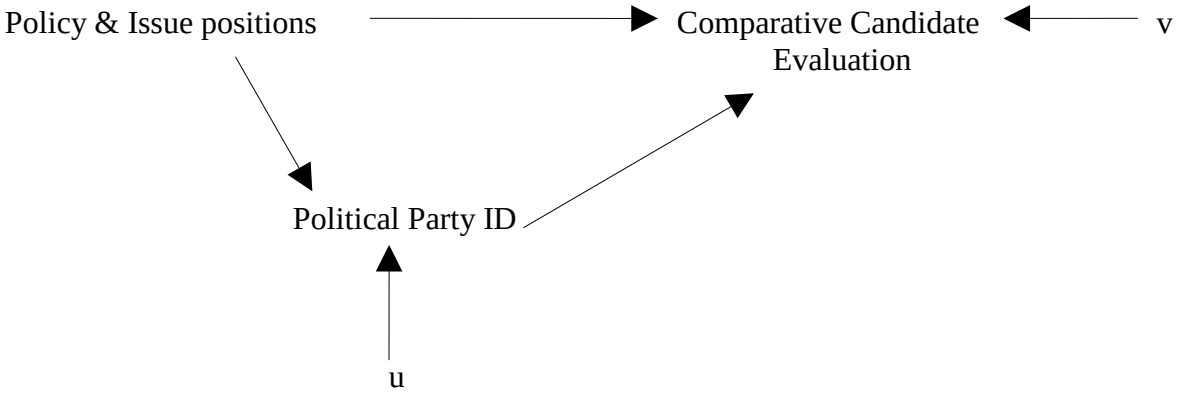
- All pairs of error (or disturbance) terms in the model are assumed to be uncorrelated.

- ε_j will be uncorrelated with all explanatory variables in the equation containing ε_j. In the above, u is uncorrelated with X1; v is uncorrelated with X1 and X2; and w is uncorrelated with X1, X2 and X3. (The disturbances can and generally will be correlated with X's that appear later in the model, e.g. u affects X2 which in turn affects X3, so u and X3 are correlated: u is an indirect cause of X3.)

- Let L = # of variables in a model (in this case 4). Recall that, for L variables, the number of unique variances and covariances = (L*[L+1]/2). So, in the above model, there are 10 unique variances and covariances. Note too that in the above, there are 10 structural parameters: 1 exogenous variance, 6 betas, and three disturbance variances. The above model is *just-identified*. If there were fewer structural parameters than there were covariances (e.g. if one or more of the betas = 0) the model would be *over-identified*. The optional Appendix A of this handout discusses identification further and how it provides an alternative view of hypothesis testing.

An advantage of recursive models is that they are easy to estimate. All recursive models are identified. OLS regression can be used to obtain unbiased estimates of the model's coefficients.

Unfortunately, in many situations the assumptions of recursive models are not realistic. Consider the following model:



According to this model, policy and issue positions affect an individual's party affiliation. Each of these in turn affects how candidates are evaluated. The assumptions of the model may not be reasonable.

- While party id may influence evaluation of candidates, it may also be the case that candidate evaluations affect party id, e.g. if you like a candidate, you may be more likely to identify with that candidate's party.

- Similarly, it is also possible that, when formulating positions on issues, citizens take cues from the political leaders and parties they support. For example, if you like George Bush and/or the Republican party, you may be more likely to favor the foreign policy positions taken by the Bush administration.

- We may also question the assumption that u and v (the disturbances) are uncorrelated. For this assumption to be reasonable, we have to believe that the factors that influence an individual's party id but have not been explicitly brought into the model are uncorrelated with the factors that influence an individual's candidate evaluations but are not explicitly in the model.

  For example, Page and Jones (1979) argue that "partisan voting history" (i.e. the degree of consistency in the individual's support for a single party in previous presidential elections) affects both (current) party id and candidate evaluation. If they are correct, then partisan voting history is reflected in the error terms for both endogenous variables. Hence, we expect u and v to be correlated, making the assumptions of the recursive model inappropriate. [NOTE: This basically says that omitted variable bias can result in violation of the assumptions that the error terms are uncorrelated.]

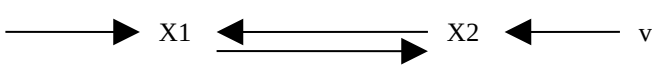
- Measurement error can also produce correlated error terms. To the extent that similar measuring devices are used to measure several endogenous variables in a model, any systematic errors produced by the measuring device will tend to be present in a similar

fashion in each of the variables, thus resulting in correlated error terms. In the present example, (1) similarities in the phrasing of questions used to measure several variables (2) the nature of the interviewer (3) similar errors in the coding of responses from question to question, and (4) other survey characteristics could result in the error terms for the equations being mutually correlated.

In sum, there is often considerable reason for doubting that the strict assumptions required for a recursive model are appropriate. Unless one is convinced that (1) causation among the variables is strictly unidirectional, and (2) the factors constituting the error terms in the model are fundamentally different for each equation, Berry says that recursive models should not be used. Instead more realistic nonrecursive models should be estimated. (Alas, this is easier said than done.)

What harms result if these problems exist but are not taken into account? *If a recursive model is employed when the assumptions required are violated and if OLS regression is used to estimate the coefficients of the model, the resulting estimates will be biased and inconsistent and, thus, will give an inaccurate assessment of the nature of the magnitude of the causal effects.*

Optional Proof: To see this, consider the following simple nonrecursive model:



Note that v affects X2 which in turn affects X1; ergo, v is correlated with X1. Similarly, u is correlated with X2. The X2 structural equation is

$X2 = \beta_{21}X1 + v$

If we multiply both sides by X1 and take expectations, we get

$\sigma_{12} = \beta_{21}\,\sigma_{11} + \sigma_{1v}$

If we subtract σ_{1v} from both sides and then divide both sides by σ_{11}, we get

$$\beta_{21} = \frac{\sigma_{12} - \sigma_{1v}}{\sigma_{11}} = \frac{\sigma_{12}}{\sigma_{11}} - \frac{\sigma_{1v}}{\sigma_{11}}$$

Recall, however, that in a sample using OLS regression,

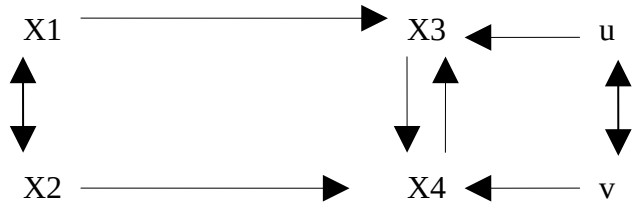
$$b_{21} = \frac{s_{12}}{s_{11}}, \text{ and } E(b_1) = \frac{\sigma_{12}}{\sigma_{11}}$$

Ergo, b_{21} will be biased by σ_{1v}/σ_{11}.

This does *not* mean, however, that we should specify models in which there are reciprocal influences between every variable (i.e. a *fully nonrecursive model*.) To be useful in empirical research, a model cannot be fully nonrecursive. Typically, some of the parameters of a nonrecursive model must be assumed to be zero, for reasons we will explain later.

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

We are going to focus on one type of non-recursive model, a model in which there is reciprocal causation. Consider the following:



In this model, there are reciprocal effects between X3 and X4. The residuals, u and v, are also correlated.

Note that there are 10 variances and covariances among the 4 X's. In the model above, there are

2 exogenous variances
1 exogenous covariance
4 structural coefficients
2 residual variances
1 residual covariance

i.e. 10 population parameters account for the 10 population variances and covariances. In this case, the model is just-identified. (Note, however, that if X1 affected X4 or X2 affected X3, there would be more parameters than there were variances and covariances, and the model would be *under-identified* – and impossible to completely estimate. We'll return to this shortly.)

There are various ways of estimating this nonrecursive model (e.g. Indirect least squares, instrumental variables, indirect least squares, LISREL models). For now, I will focus on a technique called 2 stage least squares (2SLS).

Optional: 2SLS is a *limited information technique*. Limited information techniques estimate the parameters of a nonrecursive model one equation at a time. Conversely, *full information* techniques estimate the parameters for all equations in a model simultaneously. (LISREL is a full information maximum likelihood technique.) Full information techniques produce more efficient parameter estimates, i.e. standard errors are smaller. On the other hand, full-information estimators may be more sensitive to errors in model estimation, as biases resulting from specification error in one equation tend to be transmitted through parameter estimators for all the equations in the model.

All estimation techniques require that equations be identified (which we'll return to later). 2SLS also can estimate the parameters of over-identified equations.

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 obtained to construct instrumental variables

$\hat{X}_3, \hat{X}_4,$ where

$$\hat{X}_3 = b_{31}^* X_1 + b_{32}^* X_2$$

$$\hat{X}_4 = b_{41}^* X_1 + b_{42}^* X_2$$

Note that these instrumental variables 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 the instrumental variables are computed from X1 and X2, the instrumental variables will *not* be correlated with u and v.

Optional. For example,

$$E(\hat{X}_3 v) = E[(b_{31}^* X_1 + b_{32}^* X_2)v] = b_{31}^* E(X_1 v) + b_{32}^* E(X_2 v) = 0$$

Put differently, we regress each endogenous variable on all variables in the model assumed to be uncorrelated with the model's error term. In a more complicated model, you would also regress each endogenous variable on any other predetermined endogenous variables, where a predetermined variable is not directly or indirectly affected by the dependent variable.

- 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 instrumental variable. 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.

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

The standard errors of 2SLS estimators are partially a function of the degree to which the instrumental variables created in the first stage are similar to the endogenous variables they replace. Ceterus Paribus, the higher the correlation between the instrumental 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 IVs in the first stage regressions is because we want to construct instrumental 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. Indeed, this is the way it used to be done, and Appendix C describes the procedure. But, the approach is cumbersome, and has the added disadvantage that the standardized parameter estimates, standard errors and R^2 values are all incorrect and have to be adjusted by hand. Fortunately, SPSS and other packages will now do 2SLS as a one step procedure, avoiding the problems of the 2 step

OLS approach. (Nonetheless, you may want to look at the "old" approach so you better understand the underlying logic.)

Here is a hypothetical example. We estimate the model using 2SLS. We then compare the results with an OLS regression which ignores the fact that the model is nonrecursive.

X4 Dependent

2 Stage Least Squares (Correct)

```
-> TSET NEWVAR=NONE .
-> 2SLS x4  WITH x3 x2
->   /INSTRUMENTS x1 x2
->   /CONSTANT.

MODEL:  MOD_21.

Equation number:    1

Dependent variable.. X4

Multiple R           .63580
R Square             .40424
Adjusted R Square    .40184
Standard Error      4.43898

            Analysis of Variance:
                DF   Sum of Squares      Mean Square
Regression       2        6644.9822        3322.4911
Residuals      497        9793.1747          19.7046

F =     168.61520       Signif F =  .0000

------------------ Variables in the Equation ------------------

Variable             B        SE B       Beta          T  Sig T

X2              .416696     .022901    .786907    18.196  .0000
X3              .643601     .065129    .506152     9.882  .0000
(Constant)    -1.859593    1.091455               -1.704  .0890
```

OLS (Wrong)

```
-> REGRESSION
->   /DEPENDENT x4
->   /METHOD=ENTER x2 x3 .



* * * *   M U L T I P L E   R E G R E S S I O N   * * * *

Equation Number 1    Dependent Variable..   X4


Multiple R           .61486
R Square             .37805
Adjusted R Square    .37555
Standard Error      3.99987

Analysis of Variance
                    DF      Sum of Squares      Mean Square
Regression           2          4833.29721       2416.64860
Residual           497          7951.50279         15.99900

F =     151.04998       Signif F =  .0000

------------------ Variables in the Equation ------------------

Variable             B        SE B       Beta          T  Sig T

X2              .340589     .020031    .643183    17.003  .0000
X3              .127549     .048099    .100310     2.652  .0083
(Constant)     6.193688     .831805               7.446  .0000
```

Note that the estimated effect of X3 differs greatly between the "correct" and "incorrect" approaches. Similarly, for X3

X3 Dependent

2 Stage Least Squares (Correct)

```
-> 2SLS x3  WITH x4 x1
->   /INSTRUMENTS x1 x2
->   /CONSTANT.

MODEL:  MOD_2.


Equation number:    1

Dependent variable.. X3

Multiple R           .88411
R Square             .78165
Adjusted R Square    .78077
Standard Error      1.77997

            Analysis of Variance:
                DF   Sum of Squares      Mean Square
Regression       2        5636.9814        2818.4907
Residuals      497        1574.6365           3.1683

F =     889.59569       Signif F =  .0000

------------------ Variables in the Equation ------------------

Variable             B        SE B       Beta          T  Sig T

X1              .405232     .009696    .913490    41.794  .0000
X4             -.275834     .023842   -.350738   -11.569  .0000
(Constant)     5.627888     .336290               16.735  .0000
```

OLS (Wrong)

```
-> REGRESSION
->   /DEPENDENT x3
->   /METHOD=ENTER x1 x4 .

* * * *   M U L T I P L E   R E G R E S S I O N   * * * *


Equation Number 1    Dependent Variable..   X3

Block Number  1.  Method:  Enter      X1       X4

Multiple R           .89594
R Square             .80271
Adjusted R Square    .80192
Standard Error      1.77167

Analysis of Variance
                    DF      Sum of Squares      Mean Square
Regression           2          6347.20221       3173.60110
Residual           497          1559.99779          3.13883

F =    1011.07819       Signif F =  .0000

------------------ Variables in the Equation ------------------

Variable             B        SE B       Beta          T  Sig T

X1              .411005     .009234    .926503    44.511  .0000
X4             -.311186     .016370   -.395690   -19.010  .0000
(Constant)     5.945415     .297026               20.016  .0000
```

For X3, the estimated effects do not differ that much, but there are differences in the standard errors and T values.

Stata Example. Stata has various commands that will do two stage (and also three stage) least squares. Illustrated below are the `ivregress` and `reg3` commands (see Stata's help for complete details on syntax). While results are generally identical to SPSS, one difference you'll notice is that the R^2 values are different, because the two programs use different ways of computing R^2. LIMDEP, incidentally, reports the same R^2 as Stata, but gives different F values than either Stata or SPSS does. I have to admit I don't fully understand what each program is doing differently, but luckily they all give virtually identical estimates for the parameters and their standard errors, at least for this problem.

```
. use https://www3.nd.edu/~rwilliam/statafiles/nonrecur.dta, clear
. ivregress 2sls x4 x2 (x3 = x1 x2)

Instrumental variables (2SLS) regression              Number of obs =     500
                                                      Wald chi2(2)  =  339.27
                                                      Prob > chi2   =  0.0000
                                                      R-squared     =  0.2340
                                                      Root MSE      =  4.4256

------------------------------------------------------------------------------
          x4 |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
          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
------------------------------------------------------------------------------
Instrumented:  x3
Instruments:   x2 x1

. ivregress 2sls x3 x1 (x4 = x1 x2)

Instrumental variables (2SLS) regression              Number of obs =     500
                                                      Wald chi2(2)  = 1789.93
                                                      Prob > chi2   =  0.0000
                                                      R-squared     =  0.8009
                                                      Root MSE      =  1.7746

------------------------------------------------------------------------------
          x3 |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
          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
------------------------------------------------------------------------------
Instrumented:  x4
Instruments:   x1 x2
```

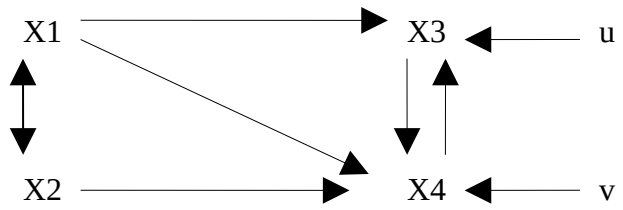
The `ivregress` command (which is the most directly analogous to the SPSS 2SLS command) is handy because you don't have to specify all the equations if you do not want to. If, on the other hand, you do, the `reg3` command is a little more convenient. By default, `reg3` does 3 stage least squares (which, among other things, makes it possible to test equality constraints across equations), although there are options for 2SLS and several other methods that are sometimes used. In this case, the 2SLS and 3SLS results are almost the same.

```
. reg3 (x4 = x3 x2) (x3 = x4 x1)

Three-stage least squares regression
------------------------------------------------------------------------------
Equation          Obs  Parms        RMSE    "R-sq"       chi2        P
------------------------------------------------------------------------------
x4                500      2    4.425647    0.2340     339.27   0.0000
x3                500      2    1.774619    0.8009    1789.93   0.0000
------------------------------------------------------------------------------

------------------------------------------------------------------------------
             |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
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
-------------+----------------------------------------------------------------
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
------------------------------------------------------------------------------
Endogenous variables:  x4 x3
Exogenous variables:   x2 x1
------------------------------------------------------------------------------
```

Using 2sls instead, you get results exactly identical to `ivregress`:

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

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

------------------------------------------------------------------------------
             |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
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
-------------+----------------------------------------------------------------
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
------------------------------------------------------------------------------
Endogenous variables:  x4 x3
Exogenous variables:   x2 x1
------------------------------------------------------------------------------
```

The Problem of Identification. Consider the following model:



The structural equations are

$$X3 = \beta_{31}X1 + \beta_{34}X4 + u$$

$$X4 = \beta_{41}X1 + \beta_{42}X2 + \beta_{43}X3 + v$$

Note that u affects X3 which in turn affects X4; hence, u and X4 are correlated. Similarly, v and X3 are correlated. Also, X4 is not predetermined in the X3 equation, and X3 is not predetermined in the X4 equation, since X3 and X4 are determined simultaneously.

If we multiply through by the predetermined variables and take expectations for the X3 variable, we get

$$\sigma_{13} = \beta_{31}\sigma_{11} + \beta_{34}\sigma_{14}$$

$$\sigma_{23} = \beta_{31}\sigma_{12} + \beta_{34}\sigma_{24}$$

There are two equations and 2 unknowns. The X3 equation is identified. It is possible to estimate the parameters for the X3 equation using 2 stage least squares.

Let us now do the same for X4:

$$\sigma_{14} = \beta_{41}\,\sigma_{11} + \beta_{42}\,\sigma_{12} + \beta_{43}\,\sigma_{13}$$

$$\sigma_{24} = \beta_{41}\,\sigma_{12} + \beta_{42}\,\sigma_{22} + \beta_{43}\,\sigma_{23}$$

Note that there are three unknowns, but only 2 equations. A unique solution for the βs is not possible. *There are an infinite number of possible solutions when there are more unknowns than there are equations.*

Optional. Here is an example from high school algebra of the problems you have when there are more unknowns than there are equations:

3A + 2B + C = 6 (1)

A + B + C = 3 (2)

If we take (2) - (1), we get

2A + B = 3 (3)

So, one solution is A = B = C = 1. However, another solution is that A = 0, B = 3, C = 0. Or, if you prefer, A = 1.5, B = 0, C = 1.5. There are an infinite number of other solutions.

To show the problem another way — suppose we attempted 2SLS on equation 4. We would estimate the model

$$X_4 = \beta_{41} X_1 + \beta_{42} X_2 + \beta_{43} \hat{X}_3 + v$$

BUT, recall that

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

That is, $\hat{X}_3$ is computed from X_1 and X_2, ergo if you try to estimate this regression there is a problem of perfect multicollinearity.

The X4 equation is therefore said to be *under-identified*. There are an infinite number of possible values for the betas that would be consistent with the observed data.

Note that the problem of identification is quite distinct from problems due to errors of sampling. *We would be unable to estimate the structural coefficients in an underidentified equation even if we knew all the population variances and covariances.*

Example.

```
******************************************************************************.
* Here is what happens if you attempt 2SLS on an underidentified equation.

* 2-Stage Least Squares the easy way -- x4 equation.
TSET NEWVAR=NONE .
2SLS x4  WITH x1 x2 x3
  /INSTRUMENTS x1 x2
  /CONSTANT.
```

Two-stage Least Squares

```
>Error # 15858
>The specified equation(s) can't be estimated because number of instrumental
>variables is too small.  For each equation, number of estimated parameters
>should not be greater than total number of instrumental variables defined
>on INSTRUMENTS subcommand.
>This command not executed.
```

The Stata equivalent is

```
. ivregress x4 x1 x2 (x3 = x1 x2)
equation not identified; must have at least as many instruments not in
the regression as there are instrumented variables
r(481);
```

In short, SPSS and Stata will not let you use 2SLS on an underidentified equation.

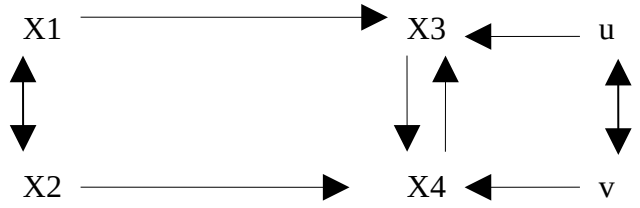
Optional. It may be helpful to compare this to what would happen if we were doing it the "hard way" using OLS. In Stage 1 we would regress X3 on X1 and X2 and compute X3HAT. In Stage 2, we would regress X4 on X1, X2, and X3HAT. This is what we would get:

```
-> REGRESSION
->   /DEPENDENT x4
->   /METHOD=ENTER x1 x2 x3hat.

------------------ Variables in the Equation ------------------

Variable             B        SE B       Beta         T  Sig T

X2              .416696     .018133    .786907    22.980  .0000
X3HAT           .643601     .051569    .427359    12.480  .0000
(Constant)    -1.859593     .864215               -2.152  .0319

------------- Variables not in the Equation -------------

Variable     Beta In  Partial  Min Toler         T  Sig T

X1           .        .          .000000      .     .
```

Note that, while SPSS provides estimates, the X1 variable is simply excluded from the equation. If included, it would have a tolerance of zero, because it is perfectly correlated with the other two variables in the equation.

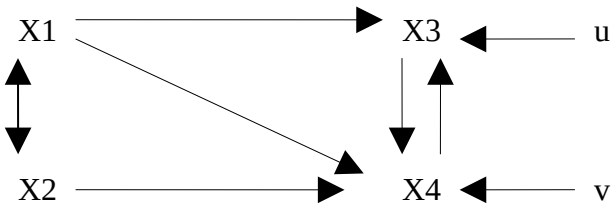
Solving the Problem of Underidentification. How, then, do we tell if an equation is underidentified—and what can we do about it if it is? Appendix B provides a more detailed explanation, but a simpler way of thinking about it is as follows:

Suppose Xi and Xj each affect each other. *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 Xi.* Appendix B discusses this point in more detail. But, in brief, 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 Xi equation, it should be the case that we think this variable affects Xi. 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. Or, if we think the added variable is actually a consequence of Xi rather than a cause of it, we produce meaningless parameter estimates. (Besides, the added variable would be correlated with the disturbance of Xi, which means it can't be used as an instrument.)

- Perhaps even more difficult, *we must believe that any added variables have indirect effects on Xj, but do not have direct effects on Xj.* That is, we have to believe that Xi is the mechanism through which the added variable affects Xj, and that once Xi is controlled for, the added variable has no direct effect on Xj. It can be quite difficult to think of such variables. This is much the same as the problem of causal ordering in recursive models. You have to be able to argue convincingly that certain logically possible direct connections between variables are, in reality, nonexistent.

If the endogenous variables in these equations are really just slightly different measures of the same thing — say, an individual's attitudes on three different but closely related issues — it is going to require a very subtle and elaborate theory indeed to produce distinct determinants of those attitudes.

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 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.

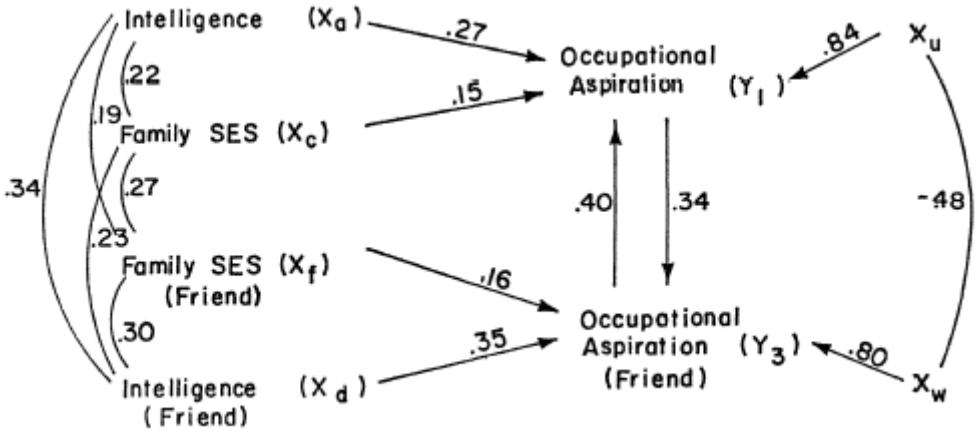
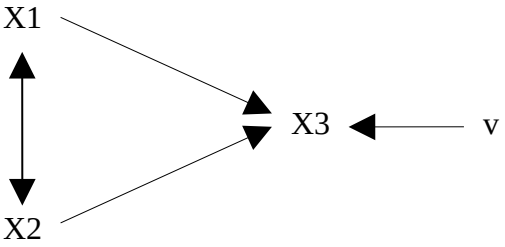


FIG. 2.—Model II

Appendix A: Identification in Recursive Models

A Just-Identified Model. Consider the model



By using Sewell Wright's rule, or else by multiplying through and taking expectations (for the variables in standardized form) we get

$$\rho_{13} = \beta_{31} + \beta_{32}\rho_{12} \qquad (1)$$

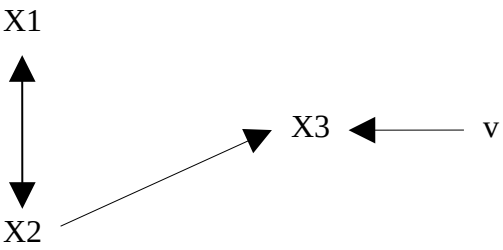
$$\rho_{23} = \beta_{31}\rho_{12} + \beta_{32} \qquad (2)$$

Note that, if the correlations (ρs) are known, we have two unknown parameters (the βs) and two equations. With two equations and two unknowns, we can solve for the unknown parameters:

$\rho_{12}\rho_{23} = \beta_{31}\rho_{12}^2 + \beta_{32}\rho_{12}$ (3)	Multiply both sides of equation (2) by ρ_{12}
$\rho_{13} - \rho_{12}\rho_{23} = \beta_{31} + \beta_{32}\rho_{12} - \beta_{31}\rho_{12}^2 - \beta_{32}\rho_{12}$ $= \beta_{31}(1-\rho_{12}^2)$ (4)	Equation (1) - Equation (3)
$\beta_{31} = \frac{\rho_{13} - \rho_{12}\rho_{23}}{1-\rho_{12}^2}$ (5)	Divide both sides of equation (4) by $(1-\rho^2_{12})$
$\beta_{32} = \frac{\rho_{23} - \rho_{12}\rho_{13}}{1-\rho_{12}^2}$ (6)	Via similar logic to the above

We say that the X3 equation is *just-identified*. More generally, in a recursive model, if all the predetermined variables affect the endogenous variable, the equation for that variable is just-identified. If all equations are just-identified, we can say that the whole model is just-identified.

An Over-Identified Model. Now consider this model:



By Sewell Wright's rule, or else by multiplying through and taking expectations (for the variables in standardized from) we get

$$\rho_{13} = \beta_{32}\rho_{12} \quad (1)$$

$$\rho_{23} = \beta_{32} \quad (2)$$

Note that there is only one unknown parameter, but two equations. Further, Equation (1) implies that

$$\beta_{32} = \frac{\rho_{13}}{\rho_{12}} \quad (3)$$

Hence, if the model is correct, then equations (2) and (3) tell us that, in the population

$$\beta_{32} = \rho_{23} = \frac{\rho_{13}}{\rho_{12}} \quad (4)$$

In this case, we say that the X3 equation is *over-identified*, and equation (4) gives the over-identifying constraint. In this case, we can say that the model is over-identified.

Of course, even if the model is true, in the sample it will probably not be the case that $r_{23} = r_{13}/r_{12}$. The question, then, is whether the difference between the observed correlations and the correlations predicted under the model are small enough to attribute to chance alone.

Hence, a test of the hypothesis

H_0: $\beta_{31} = 0$

in this case is the same as a test of the hypothesis

H_0: $\rho_{23} = \rho_{13}/\rho_{12}$

When we test whether or not paths can be eliminated from a recursive model, we are testing whether or not the overidentifying restrictions are justified.

Put another way, in an over-identified equation, the correlations implied by the model will almost certainly differ somewhat from the observed correlations. The question is whether the difference is large enough to attribute to chance or is simply due to sampling variability.

The moral, then, is that the reason we can eliminate paths from a model is because, if the paths do equal zero, the correlations will conform to constraints such as the above.

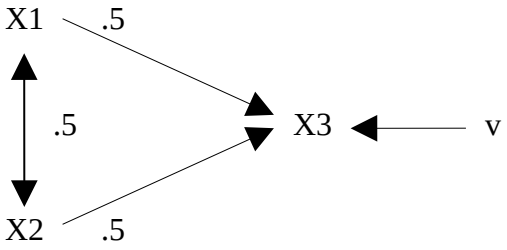
Finally, note that, within the same model, some equations can be just-identified, others can be over-identified, and yet still others can be *under-identified*.

Example. Suppose $r_{12} = .5$, $r_{13} = .75$, $r_{23} = .75$. If we estimate the just-identified model from above, we get

$$b_{31} = \frac{r_{13} - r_{12}r_{23}}{1 - r_{12}^2} = \frac{.75 - .5*.75}{1 - .5^2} = .5$$

$$b_{32} = \frac{r_{23} - r_{12}r_{13}}{1 - r_{12}^2} = \frac{.75 - .5*.75}{1 - .5^2} = .5$$

i.e.



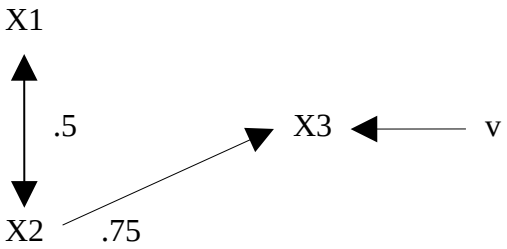
Using Sewell Wright's rule, we can now compute the correlations:

$$r_{13} = b_{31} + b_{32}r_{12} = .5 + .5*.5 = .75$$

$$r_{23} = b_{32} + b_{31}r_{12} = .5 + .5*.5 = .75$$

Ergo, the observed correlations, and the correlations implied by the model, are one and the same.

If we now estimate the over-identified model, we get $b_{32} = r_{23} = .75$, i.e.



The correlations implied by this model are

$$r_{13} = b_{32}r_{12} = .75 * .5 = .375$$

$$r_{23} = b_{32} = .75$$

In this case, the actual correlations and the correlations implied by the model are not one and the same. r_{13} is underestimated. If the difference between what the model predicts and what is actually observed is large enough, we will decide in favor of the just-identified model. If the difference is small, we will go with the over-identified model.

Appendix B: Identification in Nonrecursive Models

This supplements the discussion in the main handout. See it for additional detail.

Order Condition. The order condition is a "necessary" condition for identification. It is not a "sufficient" condition, although it tends to suffice in practice. (The "rank" condition, which I won't describe here, is a sufficient condition for identification. See Berry for a way of determining whether the rank condition is met.)

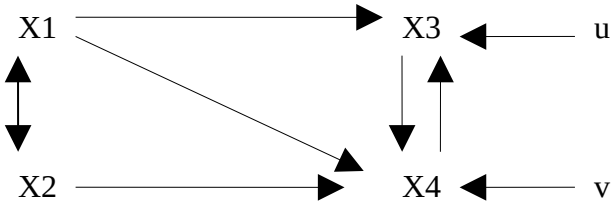
- For each equation of a model, count the number (G) of explanatory variables (variables on which the dependent variable depends directly, i.e. have causal arrows pointing directly to it). (In the SPSS 2SLS command, the explanatory variables are listed with the WITH subcommand.)

- Then count the number (H) of variables available as instrumental variables; these will include all exogenous variables in the model and any other variables that are predetermined with respect to that particular equation. (In the simple nonrecursive models presented so far, the only predetermined variables are the strictly exogenous ones.) (In the SPSS 2SLS command, the instrumental variables are listed with the INSTRUMENTS subcommand.)

- A necessary condition for identification is that $H \geq G$.

- If $H < G$, the equation is underidentified.

Consider again this model:



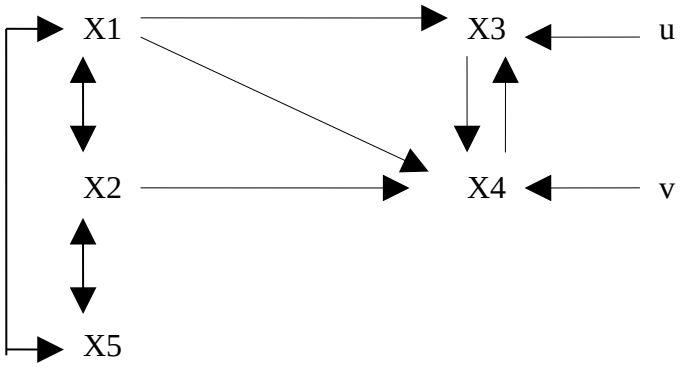
In the present example, for the X3 equation there are G = 2 explanatory variables (X1 and X4). There are H = 2 instrumental variables (X1 and X2). H = G, so the X3 equation is identified.

For the X4 equation, there are G = 3 explanatory variables (X1, X2 and X3). There are H = 2 instrumental variables (X1 and X2). H < G, so the equation is underidentified. If X1 did not affect X4, G would = 2, and you would be ok.

Another way of thinking about this: Suppose Xi and Xj each affect each other. For the Xj equation to be identified, there must be at least one predetermined variable which 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 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.

Note that, in a recursive model, the order condition is always met. This is because every explanatory variable is also an instrumental (predetermined) variable.

A Non-Solution: Dangling and Extraneous Variables. "Dangling" or "extraneous" variables don't help solve identification problems, e.g.



Here, we've seemingly added another exogenous variable. In the X4 equation G = 3 and H = 3, so the order condition is met. Alas, the X4 equation still isn't identified, because there isn't anything that affects X3 that doesn't also affect X4. (The "rank" condition is violated.) If the model is correct, when you regress X3 on X1, X2, and X5 (Stage 1 of 2SLS) the expected effect of X5 is 0 (because X5 is neither a direct nor indirect cause of X3, while X1 and X2 are.) Of course, in a sample, the effect of X5 probably would differ from zero, but only because of sampling variability (or incorrect model specification on your part).

The model would be identified if X5 affected X3 but not X4.

Similarly, it would not do any good to add an X5 that affected X1 and/or X2 but which did not directly affect X3. While X5 would have an indirect effect on X3, its direct effect would be zero once X1 and X2 were controlled.

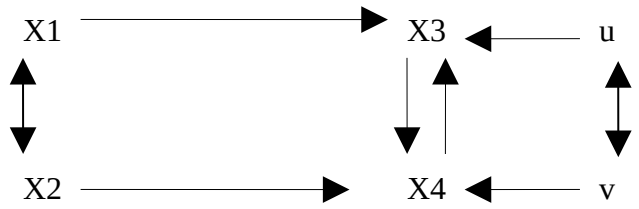
As we said before, any added variable must have a significant direct effect on X3 but not X4 in order to be useful in solving the identification problem.

Related to this is the problem of *empirical underidentification*. Suppose you added a path from X5 to X3 in your model. The model as drawn would be identified. But, if the effect of X5 on X3 is actually zero, then in reality your model is under-identified. Again, in a sample, the effect of X5 probably would differ from zero, but only because of sampling variability (or incorrect model specification on your part).

Also, it still wouldn't do us much good if the effect of X5 was nonzero but very weak. Recall that, in the example model's present form, the $\hat{X}_3$ produced by 2SLS is perfectly correlated with X1 and X2 (because it is computed from them). Suppose we now added a path from X5 to X3, but X5 only had a very small effect. The new $\hat{X}_3$ would differ only slightly from the old $\hat{X}_3$, and would be

very highly correlated with X1 and X2. Hence, adding a "weak" instrument would merely shift us from perfect multicollinearity to extreme multicollinearity. The standard errors would be very high and the parameter estimates very imprecise. Because of this, identifiability might be better viewed as varying from weak to strong rather than being an all or nothing proposition.

Appendix C: Two Stage Least Squares the Old, Hard Way



If your life depended on it (or if you just want to get a clearer picture of the underlying logic) here is how you could do 2SLS using an OLS regression routine for the above model. If done this way, (1) the metric (unstandardized) parameter estimates will be correct. (2) However, the standardized parameter estimates will be incorrect. The estimates will be attenuated because the variances of the instrumental variables created in the first stage of 2SLS are less than those of the original endogenous variables. (3) Also, the 2 stage approach produces incorrect standard errors and R^2 values. This because the standard errors and R^2 are calculated on the basis of the instrumental variables (rather than the original explanatory variables).

Stage 1: Compute X3Hat and X4Hat

```
*************************************************************************.
* 2SLS the hard way.
* Stage 1 of 2SLS -- Regress X3, X4 on X1 and X2. Save the predicted values
* for later use.
REGRESSION
  /DEPENDENT x3
  /METHOD=ENTER x1 x2
  /SAVE PRED (x3hat).
```

Regression

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients		
		B	Std. Error	Beta	t	Sig.
1	(Constant)	5.215	.378		13.798	.000
	X1	.344	.011	.776	31.890	.000
	X2	-.098	.010	-.234	-9.635	.000

a. Dependent Variable: X3

```
REGRESSION
  /DEPENDENT x4
  /METHOD=ENTER x1 x2
  /SAVE PRED (x4hat).
```

Regression

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients		
		B	Std. Error	Beta	t	Sig.
1	(Constant)	1.497	.622		2.408	.016
	X1	.221	.018	.393	12.480	.000
	X2	.354	.017	.668	21.240	.000

a. Dependent Variable: X4

Stage 2: Run the final regressions for X3 and X4

```
* 2nd step of 2SLS using OLS, X3 equation.
REGRESSION
  /DEPENDENT x3
  /METHOD=ENTER x1 x4hat.
```

Regression

Model Summary

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	.844[a]	.713	.712	2.13725

a. Predictors: (Constant), X4HAT Unstandardized Predicted Value, X1

ANOVA[b]

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	5636.981	2	2818.491	617.029	.000[a]
	Residual	2270.219	497	4.568		
	Total	7907.200	499			

a. Predictors: (Constant), X4HAT Unstandardized Predicted Value, X1

b. Dependent Variable: X3

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.
		B	Std. Error	Beta		
1	(Constant)	5.628	.404		13.938	.000
	X1	.405	.012	.913	34.808	.000
	X4HAT Unstandardized Predicted Value	-.276	.029	-.253	-9.635	.000

a. Dependent Variable: X3

Compare this with our earlier results using the 2SLS routine. The metric coefficients are the same, but estimates of R^2, standardized coefficients, and standard errors are wrong in OLS.

```
-> 2SLS x3  WITH x4 x1
->   /INSTRUMENTS x1 x2
->   /CONSTANT.

Dependent variable.. X3

Multiple R           .88411
R Square             .78165
Adjusted R Square    .78077
Standard Error      1.77997

            Analysis of Variance:
                DF   Sum of Squares      Mean Square
Regression       2        5636.9814        2818.4907
Residuals      497        1574.6365           3.1683

F =     889.59569       Signif F =  .0000

------------------ Variables in the Equation ------------------

Variable              B        SE B       Beta         T  Sig T

(Constant)     5.627888     .336290              16.735  .0000
X1              .405232     .009696    .913490    41.794  .0000
```

```
X4            -.275834     .023842   -.350738   -11.569  .0000

* 2nd step of 2SLS using OLS, X4 equation.
REGRESSION
  /DEPENDENT x4
  /METHOD=ENTER x2 x3hat.
```

Regression

Model Summary

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	.721[a]	.520	.518	3.51479

a. Predictors: (Constant), X3HAT Unstandardized Predicted Value, X2

ANOVA[b]

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	6644.982	2	3322.491	268.946	.000[a]
	Residual	6139.818	497	12.354		
	Total	12784.800	499			

a. Predictors: (Constant), X3HAT Unstandardized Predicted Value, X2

b. Dependent Variable: X4

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.
		B	Std. Error	Beta		
1	(Constant)	-1.860	.864		-2.152	.032
	X2	.417	.018	.787	22.980	.000
	X3HAT Unstandardized Predicted Value	.644	.052	.427	12.480	.000

a. Dependent Variable: X4

Compare this with our earlier results using the 2SLS routine. The metric coefficients are the same, but estimates of R^2, standardized coefficients, and standard errors are wrong in OLS.

```
-> TSET NEWVAR=NONE .
-> 2SLS x4  WITH x3 x2
->   /INSTRUMENTS x1 x2
->   /CONSTANT.

Dependent variable.. X4

Multiple R           .63580
R Square             .40424
Adjusted R Square    .40184
Standard Error      4.43898

            Analysis of Variance:
                DF   Sum of Squares      Mean Square
Regression       2        6644.9822        3322.4911
Residuals      497        9793.1747          19.7046

F =     168.61520       Signif F =  .0000

------------------ Variables in the Equation ------------------

Variable              B        SE B       Beta         T  Sig T

(Constant)    -1.859593    1.091455              -1.704  .0890
X2              .416696     .022901    .786907    18.196  .0000
X3              .643601     .065129    .506152     9.882  .0000
```

Doing the same thing in Stata,

```
. * Stage 1 of 2sls
. quietly regress x3 x1 x2
. quietly predict x3hat if e(sample)
. quietly regress x4 x1 x2
. quietly predict x4hat if e(sample)
. * Stage 2 of 2sls
. regress 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
------------------------------------------------------------------------------
```

```
. regress 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   -.1616281
------------------------------------------------------------------------------
```