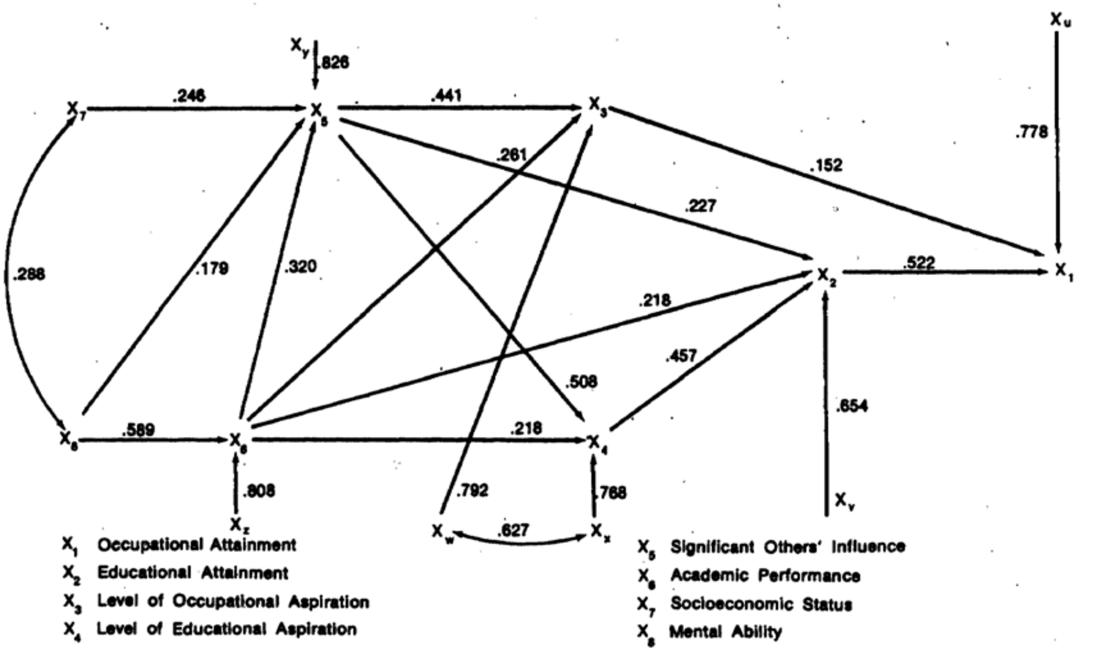
Intro to path analysis

Richard Williams, University of Notre Dame, https://www3.nd.edu/~rwilliam/

Last revised April 6, 2015

Sources. This discussion draws heavily from Otis Dudley Duncan's Introduction to Structural Equation Models.

Overview. Our theories often lead us to be interested in how a series of variables are interrelated. It is therefore often desirable to develop a system of equations, i.e. a model, which specifies all the causal linkages between variables.For example, status attainment research asks how family background, educational attainment and other variables produce socio-economic status in later life. Here is one of the early status attainment models (see Hauser, Tsai, Sewell 1983 for a discussion):



X_u
X_y
.826
X_7
.246
X_5
.441
X_3
.261
.152
.778
.227
.288
.179
.320
X_2
.522
X_1
.218
.457
.508
.654
X_8
.589
X_6
.218
X_4
.808
X_z
.792
.768
X_w
.627
X_x
X_v
X_1 Occupational Attainment
X_2 Educational Attainment
X_3 Level of Occupational Aspiration
X_4 Level of Educational Aspiration
X_5 Significant Others' Influence
X_6 Academic Performance
X_7 Socioeconomic Status
X_8 Mental Ability

Figure 2. Sewell-Haller-Ohlendorf Model of Educational and Occupational Attainment
Source: William H. Sewell, Archibald O. Haller and George W. Ohlendorf, The Educational and Early Occupational Status Attainment Process: Replication and Revision. American Sociological Review 35 (December 1970): 1023.

Among the many implications of this model are that Parents' Socio-Economic Status (X7) indirectly affects the Educational Attainment (X2) and Occupational Aspirations (X3) of children. These, in turn, directly affect children's Occupational Attainment (X1). In other words, higher parental SES helps children to become better educated and gives them higher occupational aspirations, which in turn leads to greater occupational achievement. Our earlier discussion of the Logic of Causal Order, combined with the current discussion of Path Analysis, can help us better understand how models such as the above work.

Review of key lessons from the logic of causal order. In the logic of causal order, we learned that the correlation between two variables says little about the causal relationship between them. This is because the correlation between two variables can be due to

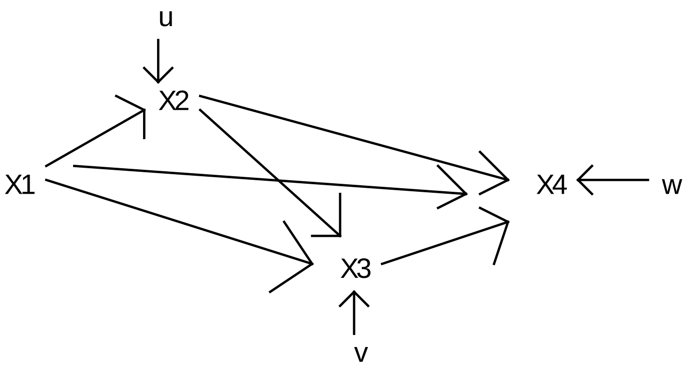
- the direct effect of one variable on another
- indirect effects; one variable affects another variable which in turn affects a third
- common causes, e.g. X affects both Y and Z. This is spurious association
- correlated causes, e.g. X is a cause of Z and X is correlated with Y
- reciprocal causation; each variable is a cause of the other

Hence, a correlation can reflect many non-causal influences. Further, a correlation can't tell you anything about the direction of causality.

At the same time, only looking at the direct effect of one variable on another may also not be optimal. Direct effects tell you how a 1 unit change in X will affect Y, holding all other variables constant. However, it may be that other variables are not likely to remain constant if X changes, e.g. a change in X can produce a change in Z which in turn produces a change in Y. Put another way, both the direct and indirect effects of X on Y must be considered if we want to know what effect a change in X will have on Y, i.e. we want to know the *total effects* (direct + indirect).

We have done all this conceptually. Now, we will see how, using path analysis, this is done mathematically and statistically. We will show how the correlation between two variables can be *decomposed* into its component parts, i.e. we will show how much of a correlation is due to direct effects, indirect effects, common causes and correlated causes. We will further show how each of the *structural effects* in a model affects the correlations in the model.

Path analysis terminology. Consider the following diagram:



In this diagram,

- X1 is an *exogenous variable*. Exogenous variables are those variables whose causes are not explicitly represented in the model. Exogenous variables are causally prior to all dependent variables in the model. There is no causal ordering of the exogenous variables. There can be more than one exogenous variable in a model. For example, if there was a 2-headed arrow linking X1 and X2 instead of a 1-headed arrow, then X1 and X2 would both be exogenous.
- Conversely, X2, X3, and X4 are *endogenous* variables. The causes of endogenous variables are specified in the model.
- Exogenous variables must always be independent variables. However, endogenous variables can be either dependent or independent. For example, X1 is a cause of X2, but X2 is itself a cause of X3 and X4.

- u, v, and w are *disturbances*, or, if you prefer, the residual terms. Many notations are used for disturbances; indeed, sometimes no notation is used at all, there is just an arrow coming in from out of nowhere. ε_2, ε_3, and ε_4 would also be a good notation, given our past practices.
- The one way arrows represent the direct causal effects in the model, also known as the *structural effects*. Sometimes, the names for these effects are specifically labeled, but other times they are left implicit. The *structural equations* in the above diagram can be written as

$$X_2 = \beta_{21} X_1 + u$$

$$X_3 = \beta_{31} X_1 + \beta_{32} X_2 + v$$

$$X_4 = \beta_{41} X_1 + \beta_{42} X_2 + \beta_{43} X_3 + w$$

- Note that we use 2 subscripts for each structural effect. The first subscript stands for the DV, the second stands for the IV. When there are multiple equations, this kind of notation is necessary to keep things straight. Note, too, that intercepts are not included. Discussions of path analysis are simplified by assuming that all variables are "centered," i.e. the mean of the variable has been subtracted from each case. Finally, note that the paths linking the disturbances to their respective variables are set equal to 1.
- In the above example, each DV was affected by all the other *predetermined variables,* i.e. those variables which are causally prior to it. We refer to such a model as being *fully recursive,* for reasons we will explain later. There is no requirement that each DV be affected by all the predetermined variables, of course. For example, β_{43} could equal zero, in which case that path would be deleted from the model. Indeed, it is fairly easy to include paths in a model; the theoretically difficult part is deciding which paths to leave out.

Determining correlations and coefficients in a path model using standardized variables.

We will now start to examine the mathematics behind a path model. For convenience, WE WILL ASSUME THAT ALL VARIABLES HAVE A MEAN OF 0 AND A VARIANCE OF 1, i.e. are standardized. This makes the math easier, and it is easy enough later on to go back to unstandardized variables. Recall that, when variables are standardized,

$E(X_1^2) = V(X_1) = 1$,

$E(X_1X_2) = COV(X_1,X_2) = \rho_{12}$ (where ρ_{12} is the population counterpart to the sample estimate r_{12})

Also, we assume (at least for now) that the disturbance in an equation is uncorrelated with any of the IVs in the equation. (Note, however, that the disturbance in each equation has a nonzero correlation with the dependent variable in that equation and (in general) with the dependent variable in each "later" equation.)

Keeping the above in mind, if we know the structural parameters, it is fairly easy to compute the underlying correlations. Perhaps more importantly, it is possible to decompose the correlation between two variables into the sources of association noted above, e.g. correlation due to direct effects, correlation due to indirect effects, etc. And, of course, if we know the correlations, we can compute the structural parameters, although this is somewhat harder to do by hand.

There are a couple of ways of doing this. The normal equations approach is more mathematical; while perhaps less intuitive, it is less prone to mistakes. second, Sewell Wright's rule, is very diagram-oriented and is perhaps more intuitive to most people once you understand it. I find that using both together is often helpful. (Both approaches are probably best learned via examples, so in class I will probably just skip to the examples and then let you re-read the following explanations on your own).

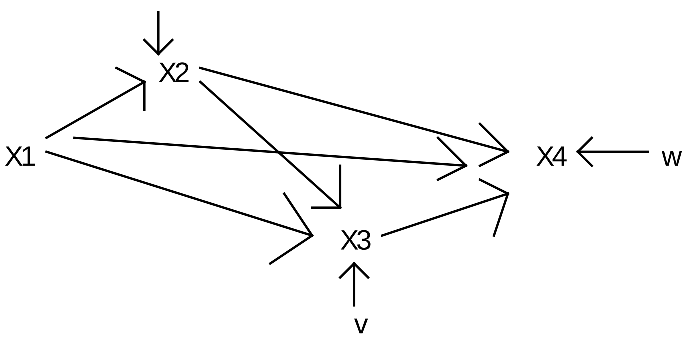
Normal equations. To get the normal equations, each structural equation is multiplied by its predetermined variables, and then expectations are taken. If the structural parameters are known, simple algebra then yields the correlations. We'll show how to use normal equations in the more complicated example.

Sewell Wright's multiplication rule: To find the correlation between X_h and X_j, where X_j appears "later" in the model,

- begin at X_j and read *back* to X_h along each distinct direct and indirect (compound) path, forming the product of the coefficients along that path. (This will give you the correlation between X_j and X_h that is due to the direct and indirect effects of X_h on X_j)
- After reading back, read *forward* (if necessary), but only one reversal from back to forward is permitted. (This will give you correlation that is due to common causes.)
- A double-headed arrow may be read either forward or backward, but you can only pass through 1 double-headed arrow on each transit. (This will give you correlation due to correlated causes)
- If you pass through a variable, you may not return to it on that transit.
- Sum the products obtained for all the linkages between X_j and X_h. (The main trick to using Wright's rule is to make sure you don't miss any linkages, count linkages twice, or make illegal double reversals.) This will give you the total correlation between the 2 variables.

To illustrate path analysis principles, we'll first go over a generic and complicated example. We'll then present a fairly simple substantive (albeit hypothetical) example similar to what we've discussed before.

Generic, Complicated Example (pretty much stolen from Duncan). We will illustrate both the Wright rule and the use of normal equations for each of the 3 structural equations in the model presented earlier:



(1) X2. For X2, the structural equation is

$$X_2 = \beta_{21} X_1 + u$$

The only predetermined variable is X1. Hence, if we multiply both sides of the above equation by X1 and then take expectations, we get the normal equation

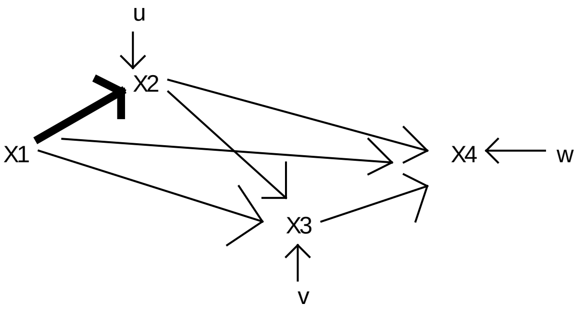
$$E(X_1 X_2) = \beta_{21} E(X_1^2) + E(X_1 u) =$$
$$\rho_{21} = \beta_{21}$$

NOTE: How did we get from the structural equation to the normal equation? First, we multiplied both sides of the structural equation by X1, and then we took the expectations of both sides, i.e.

$$X_2 = \beta_{21} X_1 + u =>$$
$$X_1 X_2 = \beta_{21} X_1{}^2 + X_1 u =>$$
$$E(X_1 X_2) = \beta_{21} E(X_1^2) + E(X_1 u) =$$
$$\rho_{21} = \beta_{21}$$

Again, remember that when variables are standardized, $E(X_1^2) = 1$ and $E(X_1X_2) = \rho_{12}$ (where ρ_{12} is the population counterpart to the sample estimate r_{12}). Also remember that we are assuming that the disturbance in an equation is uncorrelated with any of the IVs in the equation, ergo $E(X_1u) = 0$.

Hence, as we have seen before, in a bivariate regression, the correlation is the same as the standardized regression coefficient. Also, all of the correlation between X1 and X2 is causal.



SW Rule: Go back from X2 to X1.

(2) X3. For X3, the structural equation is

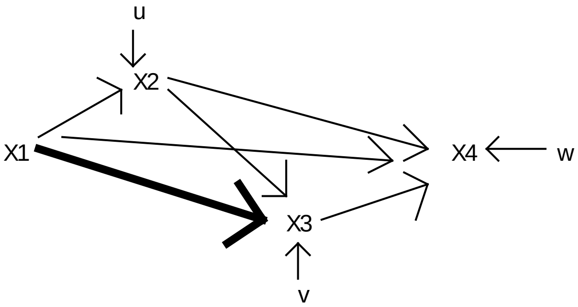
$$X_3 = \beta_{31} X_1 + \beta_{32} X_2 + v$$

There are two predetermined variables, X1 and X2. Taking each in turn, the normal equations are

$$E(X_1 X_3) = \beta_{31} E(X_1^2) + \beta_{32} E(X_1 X_2) + E(X_1 v) =$$
$$\rho_{13} = \beta_{31} + \beta_{32}\rho_{12}$$
$$= \beta_{31} + \beta_{32}\beta_{21}$$

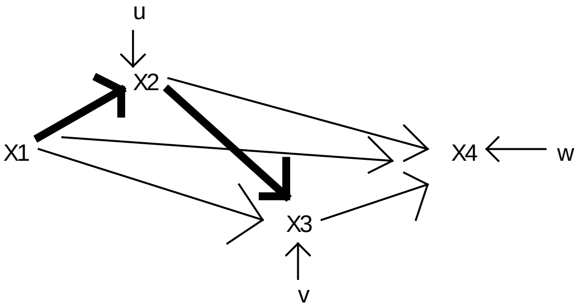
(Remember that $\beta_{21} = \rho_{12}$). As the above makes clear, there are two sources of correlation between X1 and X3:

(a) There is a direct effect of X1 on X3 (represented in β31)



SW Rule: Go back from X3 to X1.

(b) An indirect effect of X1 operating through X2 (reflected by β32β21). All of the association between X1 and X3 is causal.



SW Rule: Go back from X3 to X2, and then back from X2 to X1.

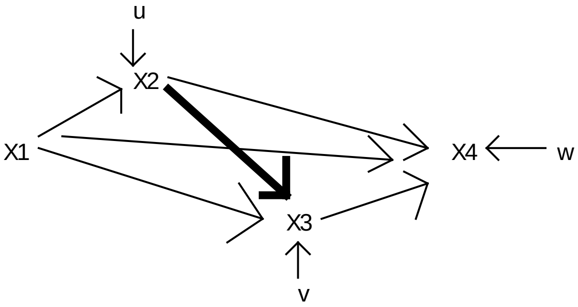
NOTE: Recall that the sum of a variable's direct effect and its indirect effects is known as its *total effect*. So, in this case, the total effect of X1 on X3 is $\beta_{31} + \beta_{32}\beta_{21}$.

Doing the same thing for X2 and X3, we get

$$E(X_2X_3) = \beta_{31}E(X_1X_2) + \beta_{32}E(X_2^2) + E(X_2v) =$$
$$\rho_{23} = \beta_{31}\rho_{12} + \beta_{32}$$
$$= \beta_{31}\beta_{21} + \beta_{32}$$

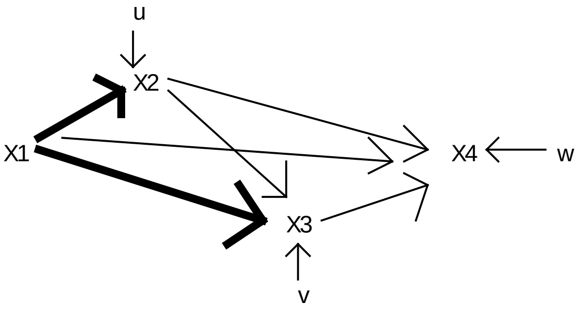
Again, as the above makes clear, there are two sources of correlation between X2 and X3:

(a) There is a direct effect of X2 on X3 (represented in β32).



SW Rule: Go back from X3 to X2.

(b) But, there is also correlation due to a common cause, X1 (reflected by $\beta_{31}\beta_{21}$). Hence, part of the correlation between X2 and X3 is spurious.



SW Rule: Go back from X3 to X1, go forward from X1 to X2.

(3) X4. For X4, the predetermined variables are X1, X2, and X3. The structural equation is

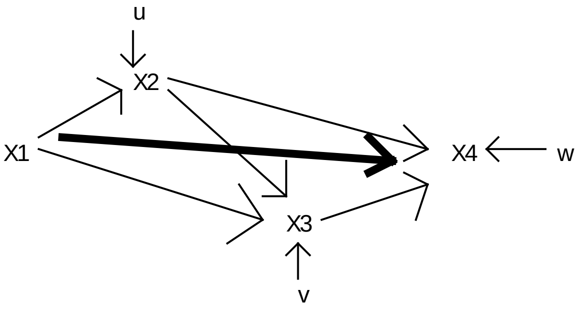
$$X_4 = \beta_{41}X_1 + \beta_{42}X_2 + \beta_{43}X_3 + w$$

The normal equations are, first, for X1,

$$\begin{aligned} E(X_1X_4) &= \beta_{41}E(X_1^2) + \beta_{42}E(X_1X_2) + \beta_{43}E(X_1X_3) + E(X_1w) = \\ \rho_{41} &= \beta_{41} + \beta_{42}\rho_{12} + \beta_{43}\rho_{13} \\ &= \beta_{41} + \beta_{42}\beta_{21} + \beta_{43}(\beta_{31} + \beta_{32}\beta_{21}) \\ &= \beta_{41} + \beta_{42}\beta_{21} + \beta_{43}\beta_{31} + \beta_{43}\beta_{32}\beta_{21} \end{aligned}$$

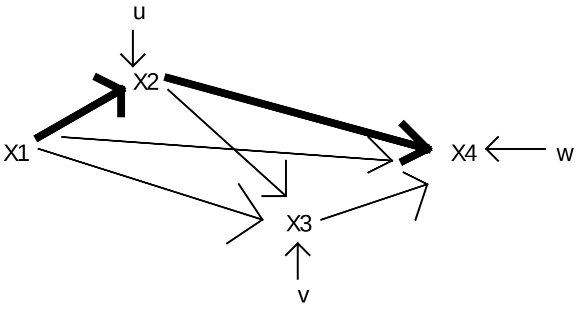
This shows there are 4 sources of association between X1 and X4:

(a) Association due to the direct effect of X1 on X4 (β_{41})



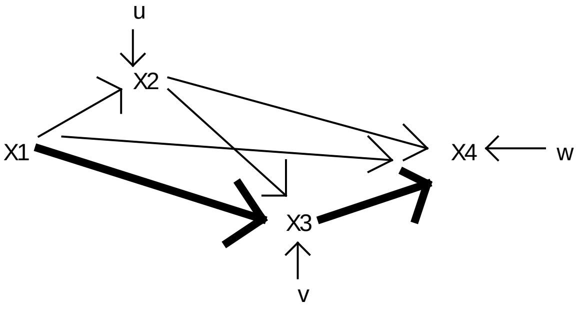
SW Rule: Go back from X4 to X1.

(b) Association due to an indirect effect: X1 affects X2 which then affects X4 ($\beta_{42}\beta_{21}$)



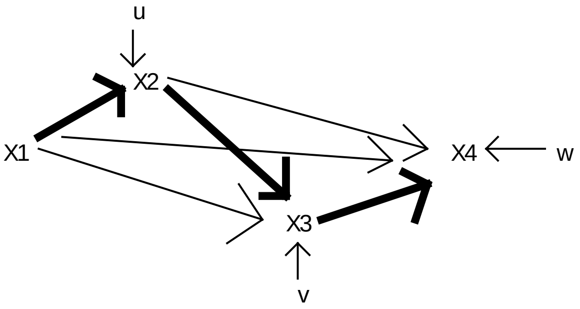
SW Rule: Go back from X4 to X2, go back from X2 to X1.

(c) Association due to another indirect effect: X1 affects X3 which then affects X4 ($\beta_{43}\beta_{31}$)



SW Rule: Go back from X4 to X3, go back from X3 to X1.

(d) Association due to yet another indirect effect: X1 affects X2, which then affects X3, which then affects X4 ($\beta_{43}\beta_{32}\beta_{21}$)



SW Rule: Go back from X4 to X3, back from X3 to X2, back from X2 to X1.

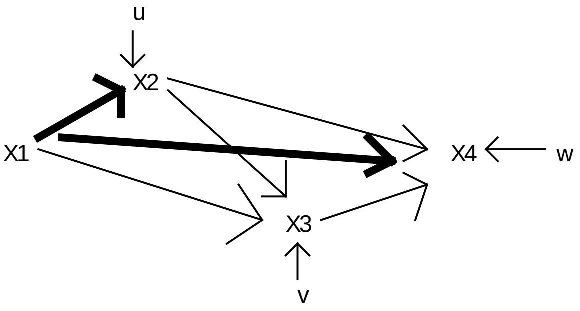
Note that you sum (b), (c) and (d) to get the total indirect effect of X1 on X4. Note too that all of the correlation between X1 and X4 is causal.

The normal equations for X2 and X4 are

$$\begin{aligned} E(X_2X_4) &= \beta_{41}E(X_2X_1) + \beta_{42}E(X_2^2) + \beta_{43}E(X_2X_3) + E(X_2w) = \\ \rho_{42} &= \beta_{41}\rho_{12} + \beta_{42} + \beta_{43}\rho_{23} \\ &= \beta_{41}\beta_{21} + \beta_{42} + \beta_{43}(\beta_{32} + \beta_{31}\beta_{21}) \\ &= \beta_{41}\beta_{21} + \beta_{42} + \beta_{43}\beta_{32} + \beta_{43}\beta_{31}\beta_{21} \end{aligned}$$

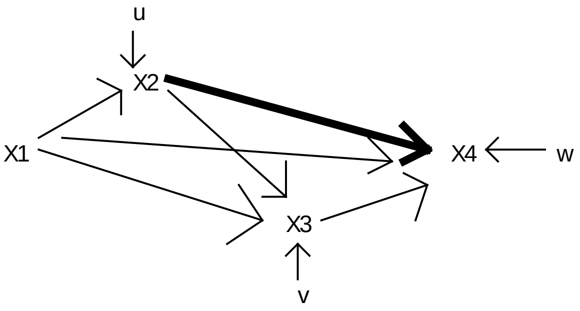
This shows there are 4 sources of association between X2 and X4:

(a) Association due to X1 being a common cause of X2 and X4 ($\beta_{41}\beta_{21}$)



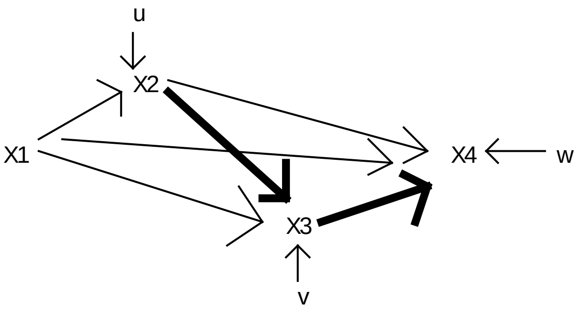
SW Rule: GO back from X4 to X1, go forward from X1 to X2.

(b) Association due to the direct effect of X2 on X4 (β_{42})



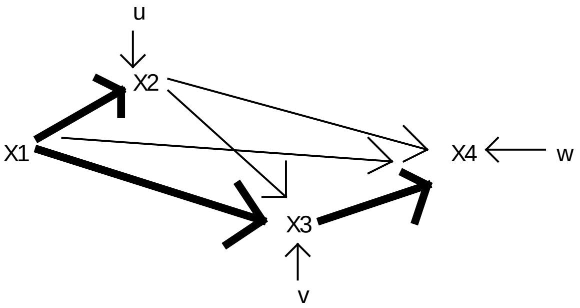
SW Rule: Go back from X4 to X2.

(c) Association due to the indirect effect of X2 affecting X3 which in turn affects X4 ($\beta_{43}\beta_{32}$)



SW Rule: Go back from X4 to X3, go back from X3 to X2.

(d) Association due to X1 being a common cause of X2 and X4: X1 directly affects X2 and indirectly affects X4 through X3 ($\beta_{43}\beta_{31}\beta_{21}$).



SW Rule: Go back from X4 to X3, back from X3 to X1, forward from X1 to X2.

Note that you sum (a) and (d) to get the correlation due to common causes. This represents spurious association, while (b) + (c) represents causal association.

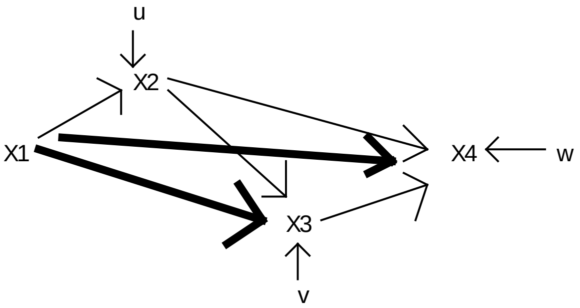
The normal equations for X3 and X4 are,

$$E(X_3X_4) = \beta_{41}E(X_3X_1) + \beta_{42}E(X_3X_2) + \beta_{43}E(X_3^2) + E(X_3w) =$$

$$\begin{aligned}\rho_{43} &= \beta_{41}\rho_{13} + \beta_{42}\rho_{23} + \beta_{43} \\ &= \beta_{41}(\beta_{31} + \beta_{32}\beta_{21}) + \beta_{42}(\beta_{32} + \beta_{31}\beta_{21}) + \beta_{43} \\ &= \beta_{41}\beta_{31} + \beta_{41}\beta_{32}\beta_{21} + \beta_{42}\beta_{32} + \beta_{42}\beta_{31}\beta_{21} + \beta_{43}\end{aligned}$$

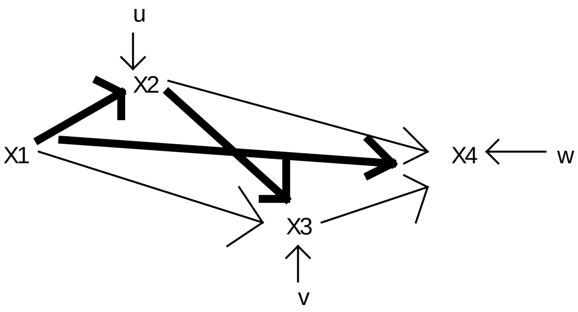
This shows there are 5 sources of association between X3 and X4:

(a) Association due to X1 being a common cause of X3 and X4 ($\beta_{41}\beta_{31}$)



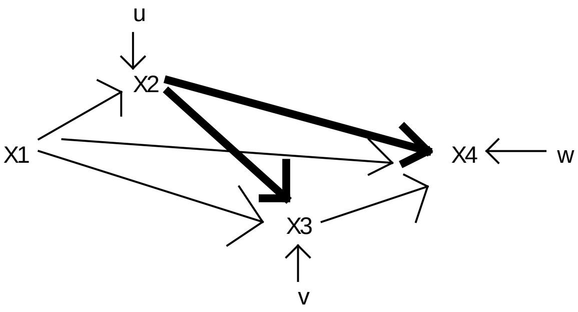
SW Rule: Go back from X4 to X1, go forward from X1 to X3.

(b) Association due to X1 being a common cause of X3 (by first affecting X2, which in turn affects X3) and X4 ($\beta_{41}\beta_{21}\beta_{32}$)



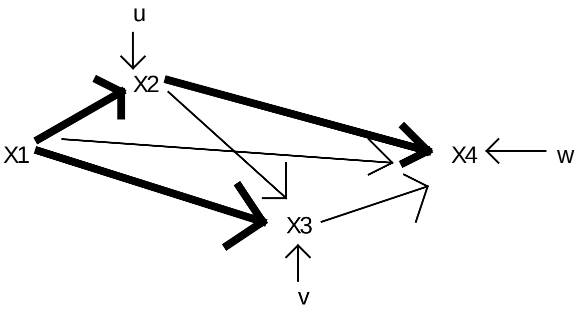
SW rule: Go back from X4 to X1, forward from X1 to X2, forward from X2 to X3.

(c) Association due to X2 being a common cause of X3 and X4 ($\beta_{42}\beta_{32}$)



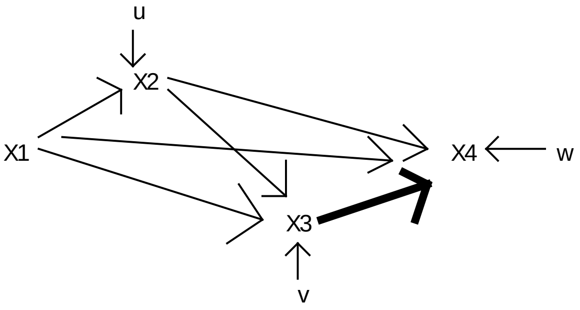
SW Rule: Back from X4 to X2, go forward from X2 to X3.

(d) Association due to X1 being a common cause of X3 and X4: X1 directly affects X3 and indirectly affects X4 through X2 ($\beta_{42}\beta_{21}\beta_{31}$).



SW Rule: Go back from X4 to X2, back from X2 to X1, forward from X1 to X3.

(e) Association due to X3 being a direct cause of X4 (β_{43})



SW Rule: Go back from X4 to X3.

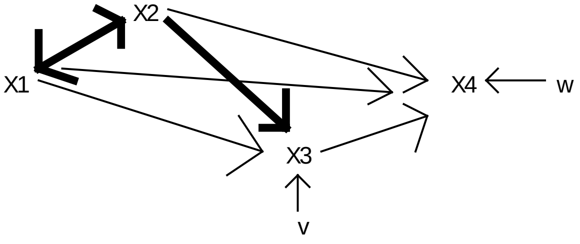
Note that you sum (a), (b) (c) and (d) to get the correlation due to common causes. This is the spurious association. There are no indirect effects of X3 on X4.

In reviewing the above, note that, *if there are no double-headed arrows in the model*

- If you go back once and then stop, it is a direct effect
- If you go back 2 or more times and never come forward, it is an indirect effect
- If you go back and later come forward, it is correlation due to a common cause

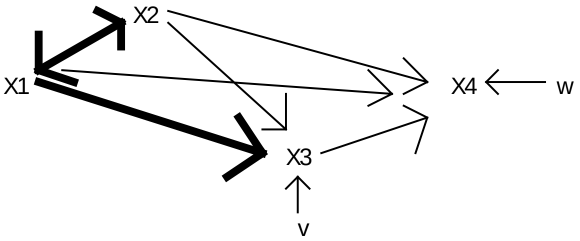
Correlated causes. Suppose that, in the above model, X1 and X2 were both exogenous, i.e. there was a double-headed arrow between them instead of a 1-way arrow. This would not have any significant effect on the math, but it would affect our interpretation of the sources of correlation. Anything involving ρ_{12} would then have to be interpreted as correlation due to correlated causes. Further, we could not always say what effect changes in X1 would have on other variables, since we wouldn't know whether changes in X1 would also produce changes in X2 (unless we have good reasons for believing that that couldn't be the case, e.g. gender and race might both be exogeneous variables in a model, but we are pretty confident that changes in one are not going to produce changes in the other.). That is, with two-headed arrows we often can't be sure what the indirect effects are, which also means that we can't be sure what the total effects are. Ergo, *the fewer 2-headed arrows in a model, the more powerful the model is in terms of the statements it makes.*

For example:



Instead of X1 and X3 being correlated because of the indirect effect of X1 affecting X2 which in turn affects X3 (which is a causal relationship) X1 and X3 are correlated because of the correlated causes of X1 and X2 (which we do not assume to be causal), i.e. X1 is correlated with a cause of X3.

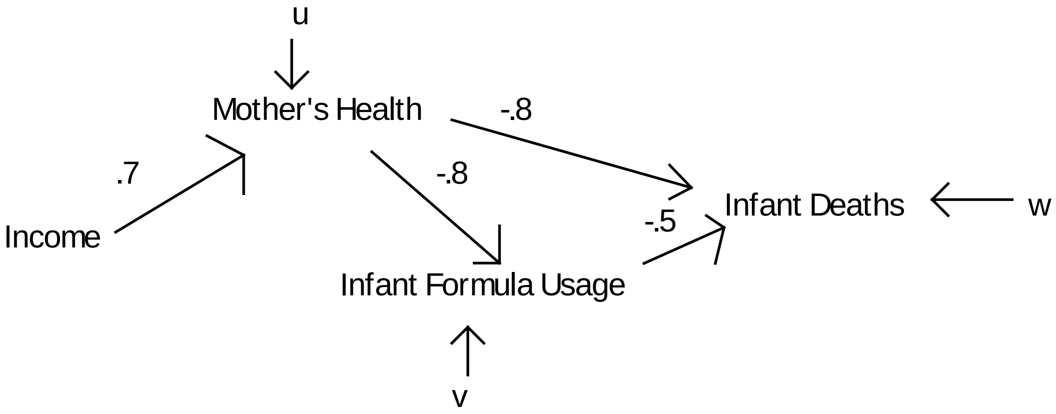
Or,



Instead of X2 and X3 being correlated because they share a common cause, they are correlated because of a correlated cause, i.e. X1 is a cause of X3 and X2 is correlated with X1.

SUBSTANTIVE HYPOTHETICAL EXAMPLE (Adapted From the 1995 Soc 593 Exam 2):

A demographer believes that the following model describes the relationship between Income, Health of the Mother, Use of Infant formula, and Infant deaths. All variables are in standardized form. The hypothesized value of each path is included in the diagram.



a. Write out the structural equation for each endogenous variable.

$$MH = \beta_{MH,Inc} Income + u = .7 * Income + u$$

$$IF = \beta_{If,MH} MH + v = -.8 * MH + v$$

$$ID = \beta_{ID,MH} * MH + \beta_{ID,IF} * IF + w = -.8 * MH - .5 * IF + w$$

b. Determine the complete correlation matrix. (Remember, variables are standardized. You can use either normal equations or Sewell Wright, but you might want to use both as a double-check.)

Correlation	Sewell-Wright Approach
$r_{mh,inc}$ = .7	Go back from Mother’s health to Income. (Direct effect of Income on MH)
$r_{if,MH}$ = -.8	Go back from IF to MH. (Direct effect of MH on IF)
$r_{IF,Inc}$ = -.8 * .7 = -.56	Go back from IF to MH, then back from MH to income. (Indirect effect of Income – Income affects mother’s health which in turn affects Infant formula usage)
$r_{id,IF}$ = -.5 + -.8*-.8 = .14	Go back from ID to IF. (Direct effect of Infant formula on infant deaths) Then, go back from ID to MH, then go forward from MH to IF. (Mother’s health is a common cause of both Infant formula usage and infant deaths) Note that, even though the direct effect of infant formula usage on infant deaths is negative (which means that using formula reduces infant deaths) the correlation between infant formula usage and infant deaths is positive (which means that those who use formula are more likely to experience infant deaths). We discuss this further below.
$r_{id,MH}$ = -.8 + -.8*-.5 = -.4	Go back from ID to MH. (Direct effect of Mother’s Health on Infant deaths) Then, go back from ID to IF to MH. (Indirect effect of Mother’s health on infant deaths – Mother’s health affects infant formula usage which in turn affects infant deaths)
$r_{id,INC}$ = -.8*.7 + -.5*-.8*.7 = -.28	Go back from Infant Death to Mother’s Health, then back to Income. (Income is an indirect cause of Infant deaths – Income affects mother’s health which in turn affects infant deaths.) Then go back from Infant deaths, then back to Mother’s Health, then back to Income. (Income is yet again an indirect cause – Income affects Mother’s Health, which affects Infant Formula Usage, which affects Infant Deaths.)

c. Decompose the correlation between Infant deaths and Usage of Infant formula into

- Correlation due to direct effects

  -.5 (see path from IF to ID)

- Correlation due to common causes

  -.8 * -.8 = .64 (Mother’s health is a cause of both IF and ID)

d. Suppose the above model is correct, but instead the researcher believed in and estimated the following model:

Infant Formula Usage → Infant Deaths ← w

What conclusions would the researcher likely draw? Why would he make these mistakes? Discuss the consequences of this mis-specification.

The correlation between IF and ID is positive, hence, if the above model was estimated, the expected value of the coefficient would be .14. This would imply that infant formula usage increases infant deaths, when in reality the correct model shows that it decreases them. The correlation is positive because of the common cause of Mother’s health: less healthy mothers are more likely to use infant formula, and they are also more likely to have higher infant death rates. Belief in the above model could lead to a reduction in infant formula usage, which would have exactly the opposite effect of what was intended.

Appendix: Basic Path Analysis with Stata

We have been doing things a bit backwards here. We have been starting with the coefficients, and then figured out what the correlations must be. Normally, of course, we start with the data/correlations and then estimate the coefficients.

Nonetheless, we can use Stata to verify we have calculated the correlations correctly. Just give Stata the correlations we computed by hand and then use one of the methods below to estimate the various regressions. If we've done everything right, the regression parameters should come out the same as in the path diagram. Remember, this is easier if you use the "input matrix by hand" submenu. (Click Data/ Matrices / Input matrix by hand.)

```
. matrix input Corr = (1,.7,-.56,-.28\.7,1,-.80,-.40\-.56,-.80,1,.14\-.28,-.40,.14,1)
. matrix input SDs = (1,1,1,1)
. matrix input Means = (0,0,0,0)
. corr2data income mhealth formula death, corr(Corr) mean(Means) sd(SDs) n(100)
(obs 100)
```

There are now at least three ways to estimate the path models (or at least, the simple models we are estimating here; approach 2, the sem commands, is probably best for more complicated models.)

I. Estimate separate regressions for each dependent variable.

```
. reg  mhealth income

      Source |       SS       df       MS              Number of obs =     100
-------------+------------------------------           F(  1,    98) =   94.16
       Model |  48.5099991     1  48.5099991           Prob > F      =  0.0000
    Residual |  50.4899995    98  .515204077           R-squared     =  0.4900
-------------+------------------------------           Adj R-squared =  0.4848
       Total |  98.9999987    99  .999999986           Root MSE      =  .71778

------------------------------------------------------------------------------
     mhealth |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
      income |         .7   .0721393     9.70   0.000     .5568419    .8431581
       _cons |   6.41e-10   .0717777     0.00   1.000    -.1424405    .1424405
------------------------------------------------------------------------------

. reg  formula  income mhealth

      Source |       SS       df       MS              Number of obs =     100
-------------+------------------------------           F(  2,    97) =   86.22
       Model |  63.3600001     2       31.68           Prob > F      =  0.0000
    Residual |       35.64    97  .367422681           R-squared     =  0.6400
-------------+------------------------------           Adj R-squared =  0.6326
       Total |  99.0000001    99           1           Root MSE      =  .60615

------------------------------------------------------------------------------
     formula |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
      income |   4.93e-09   .0853061     0.00   1.000    -.1693091    .1693091
     mhealth |        -.8   .0853061    -9.38   0.000    -.9693091   -.6306909
       _cons |  -2.31e-09   .0606154    -0.00   1.000    -.1203048    .1203048
------------------------------------------------------------------------------
```

```
. reg  death income mhealth formula

      Source |       SS       df       MS              Number of obs =     100
-------------+------------------------------           F(  3,    96) =   10.67
       Model |   24.749999     3  8.24999966           Prob > F      =  0.0000
    Residual |  74.2500011    96  .773437511           R-squared     =  0.2500
-------------+------------------------------           Adj R-squared =  0.2266
       Total |  99.0000001    99           1           Root MSE      =  .87945

------------------------------------------------------------------------------
       death |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
      income |   1.63e-09   .1237684     0.00   1.000    -.2456784    .2456784
     mhealth |        -.8   .1709021    -4.68   0.000    -1.139238   -.4607621
     formula |        -.5   .1473139    -3.39   0.001    -.7924158   -.2075842
       _cons |  -6.54e-09   .0879453    -0.00   1.000      -.17457      .17457
------------------------------------------------------------------------------

. * The mis-specified model
. reg  death formula

      Source |       SS       df       MS              Number of obs =     100
-------------+------------------------------           F(  1,    98) =    1.96
       Model |  1.94039993     1  1.94039993           Prob > F      =  0.1648
    Residual |  97.0596001    98  .990404083           R-squared     =  0.0196
-------------+------------------------------           Adj R-squared =  0.0096
       Total |  99.0000001    99           1           Root MSE      =  .99519

------------------------------------------------------------------------------
       death |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
     formula |        .14   .1000204     1.40   0.165    -.0584872    .3384872
       _cons |  -5.23e-09    .099519    -0.00   1.000    -.1974923    .1974923
------------------------------------------------------------------------------
```

II. The sem commands. We can also use the `sem` (Structural Equation Modeling) commands that were introduced in Stata 11. This example is pretty simple so it isn't too hard to do. Among the nice features of sem is that you can specify all the equations at once, and you can get estimates of the direct, indirect and total effects. Time permitting, we will talk about `sem` more later in the semester.

```
. sem (mhealth <- income) (formula <- income mhealth) (death <- income mhealth formula)

Endogenous variables

Observed:  mhealth formula death

Exogenous variables

Observed:  income

Fitting target model:

Iteration 0:   log likelihood = -466.43145
Iteration 1:   log likelihood = -466.43145

Structural equation model                       Number of obs      =       100
Estimation method  = ml
Log likelihood     = -466.43145
```

```
------------------------------------------------------------------------------
             |                 OIM
             |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
Structural   |
  mhealth <- |
      income |         .7   .0714143     9.80   0.000     .5600306    .8399694
       _cons |   6.41e-10   .0710563     0.00   1.000    -.1392678    .1392678
  -----------+----------------------------------------------------------------
  formula <- |
     mhealth |        -.8   .0840168    -9.52   0.000    -.9646699   -.6353301
      income |   4.93e-09   .0840168     0.00   1.000    -.1646699    .1646699
       _cons |  -2.31e-09   .0596992    -0.00   1.000    -.1170084    .1170084
  -----------+----------------------------------------------------------------
  death <-   |
     mhealth |        -.8   .1674491    -4.78   0.000    -1.128194   -.4718057
     formula |        -.5   .1443376    -3.46   0.001    -.7828964   -.2171036
      income |   1.63e-09   .1212678     0.00   1.000    -.2376805    .2376805
       _cons |  -6.54e-09   .0861684    -0.00   1.000     -.168887     .168887
-------------+----------------------------------------------------------------
Variance     |
   e.mhealth |      .5049   .0714036                      .3826725    .6661675
   e.formula |      .3564   .0504026                      .2701218    .4702359
     e.death |      .7425   .1050054                      .5627537    .9796581
------------------------------------------------------------------------------
LR test of model vs. saturated: chi2(0)   =      0.00, Prob > chi2 =      .

. * Estimate the direct, indirect, and total effects of each variable
. estat teffects

Direct effects
------------------------------------------------------------------------------
             |                 OIM
             |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
Structural   |
  mhealth <- |
      income |         .7   .0714143     9.80   0.000     .5600306    .8399694
  -----------+----------------------------------------------------------------
  formula <- |
     mhealth |        -.8   .0840168    -9.52   0.000    -.9646699   -.6353301
      income |   4.93e-09   .0840168     0.00   1.000    -.1646699    .1646699
  -----------+----------------------------------------------------------------
  death <-   |
     mhealth |        -.8   .1674491    -4.78   0.000    -1.128194   -.4718057
     formula |        -.5   .1443376    -3.46   0.001    -.7828964   -.2171036
      income |   1.63e-09   .1212678     0.00   1.000    -.2376805    .2376805
------------------------------------------------------------------------------
```

```
Indirect effects
------------------------------------------------------------------------------
             |                 OIM
             |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
Structural   |
  mhealth <- |
      income |          0  (no path)
  -----------+----------------------------------------------------------------
  formula <- |
     mhealth |          0  (no path)
      income |       -.56   .0819928    -6.83   0.000     -.720703    -.399297
  -----------+----------------------------------------------------------------
  death <-   |
     mhealth |         .4   .0420084     9.52   0.000      .317665     .482335
     formula |          0  (no path)
      income |       -.28   .0944557    -2.96   0.003    -.4651298   -.0948702
------------------------------------------------------------------------------


Total effects
------------------------------------------------------------------------------
             |                 OIM
             |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
Structural   |
  mhealth <- |
      income |         .7   .0714143     9.80   0.000     .5600306    .8399694
  -----------+----------------------------------------------------------------
  formula <- |
     mhealth |        -.8   .0840168    -9.52   0.000    -.9646699   -.6353301
      income |       -.56   .0828493    -6.76   0.000    -.7223816   -.3976184
  -----------+----------------------------------------------------------------
  death <-   |
     mhealth |        -.4   .1726381    -2.32   0.021    -.7383645   -.0616355
     formula |        -.5   .1443376    -3.46   0.001    -.7828964   -.2171036
      income |       -.28       .096    -2.92   0.004    -.4681565   -.0918435
------------------------------------------------------------------------------

. * Incorrect model
. sem death  <- formula

Endogenous variables

Observed:  death

Exogenous variables

Observed:  formula

Fitting target model:

Iteration 0:   log likelihood = -281.79294
Iteration 1:   log likelihood = -281.79294

Structural equation model                       Number of obs      =       100
Estimation method  = ml
Log likelihood     = -281.79294
```

```
------------------------------------------------------------------------------
             |                 OIM
             |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
Structural   |
  death <-   |
     formula |        .14   .0990152     1.41   0.157    -.0540661    .3340661
       _cons |  -5.23e-09   .0985188    -0.00   1.000    -.1930934    .1930934
-------------+----------------------------------------------------------------
Variance     |
     e.death |    .970596    .137263                      .7356317    1.280609
------------------------------------------------------------------------------
LR test of model vs. saturated: chi2(0)   =      0.00, Prob > chi2 =      .
```

III. UCLA's pathreg command. You can get this with the `findit` command. Again, it lets you specify all the equations at once, but doesn't offer the many additional features that `sem` does. Also, `pathreg` does not support factor variables as of March 2013.

```
. pathreg (mhealth income) (formula income mhealth) (death income mhealth formula)

------------------------------------------------------------------------------
     mhealth |      Coef.   Std. Err.      t    P>|t|                     Beta
-------------+----------------------------------------------------------------
      income |         .7   .0721393     9.70   0.000                       .7
       _cons |   6.41e-10   .0717777     0.00   1.000                        .
------------------------------------------------------------------------------
                 n = 100  R2 = 0.4900  sqrt(1 - R2) = 0.7141

------------------------------------------------------------------------------
     formula |      Coef.   Std. Err.      t    P>|t|                     Beta
-------------+----------------------------------------------------------------
      income |   4.93e-09   .0853061     0.00   1.000                 4.93e-09
     mhealth |        -.8   .0853061    -9.38   0.000                      -.8
       _cons |  -2.31e-09   .0606154    -0.00   1.000                        .
------------------------------------------------------------------------------
                 n = 100  R2 = 0.6400  sqrt(1 - R2) = 0.6000

------------------------------------------------------------------------------
       death |      Coef.   Std. Err.      t    P>|t|                     Beta
-------------+----------------------------------------------------------------
      income |   1.63e-09   .1237684     0.00   1.000                 1.63e-09
     mhealth |        -.8   .1709021    -4.68   0.000                      -.8
     formula |        -.5   .1473139    -3.39   0.001                      -.5
       _cons |  -6.54e-09   .0879453    -0.00   1.000                        .
------------------------------------------------------------------------------
                 n = 100  R2 = 0.2500  sqrt(1 - R2) = 0.8660
```