Structural Equation Models

Appendix to *An R and S-PLUS Companion to Applied Regression*

John Fox

Corrected (but not updated) 4 August 2006

1 Introduction

Structural equation models (*SEMs*), also called *simultaneous equation models*, are multivariate (i.e., multiple-equation) regression models. Unlike the more traditional multivariate linear model, however, the response variable in one regression equation in an SEM may appear as a predictor in another equation; indeed, variables in an SEM may influence one-another reciprocally, either directly or through other variables as intermediaries. These *structural equations* are meant to represent causal relationships among the variables in the model.

A cynical view of SEMs is that their popularity in the social sciences reflects the legitimacy that the models appear to lend to causal interpretation of observational data, when in fact such interpretation is no less problematic than for other kinds of regression models applied to observational data. A more charitable interpretation is that SEMs are close to the kind of informal thinking about causal relationships that is common in social-science theorizing, and that, therefore, these models facilitate translating such theories into data analysis. In economics, in contrast, structural-equation models may stem from *formal* theory.

To my knowledge, the only facility in S for fitting structural equation models is my `sem` library, which at present is available for R but not for S-PLUS. The `sem` library includes functions for estimating structural equations in observed-variables models by two-stage least squares, and for fitting general structural equation models with multinormal errors and latent variables by full-information maximum likelihood. These methods are covered (along with the associated terminology) in the subsequent sections of the appendix. As I write this appendix, the `sem` library is in a preliminary form, and the capabilities that it provides are modest compared with specialized structural equation software.

Structural equation modeling is a large subject. Relatively brief introductions may be found in Fox (1984: Ch. 4) and in Duncan (1975); Bollen (1989) is a standard book-length treatment, now slightly dated; and most general econometric texts (e.g., Greene, 1993: Ch. 20; Judge et al., 1985: Part 5) take up at least observed-variables structural equation models.

2 Observed-Variables Models and Two-Stage Least-Squares Estimation

2.1 An Example: Klein's Model

Klein's (1950) macroeconomic model of the U. S. economy often appears in econometrics texts (e.g., Greene, 1993) as a simple example of a structural equation model:

$$\begin{aligned} C_t &= \gamma_{10} + \gamma_{11}P_t + \gamma_{12}P_{t-1} + \beta_{11}(W_t^p + W_t^g) + \zeta_{1t} \\ I_t &= \gamma_{20} + \gamma_{21}P_t + \gamma_{22}P_{t-1} + \beta_{21}K_{t-1} + \zeta_{2t} \\ W_t^p &= \gamma_{30} + \gamma_{31}A_t + \beta_{31}X_t + \beta_{32}X_{t-1} + \zeta_{3t} \\ X_t &= C_t + I_t + G_t \\ P_t &= X_t - T_t - W_t^p \\ K_t &= K_{t-1} + I_t \end{aligned} \tag{1}$$

- The variables on the left-hand side of the structural equations are *endogenous variables* — that is, variables whose values are determined by the model. There is, in general, one structural equation for each endogenous variable in an SEM.[1]
- The ζ's (Greek *zeta*) are error variables, also called *structural disturbances* or *errors in equations*; they play a role analogous to the error in a single-equation regression model. It is not generally assumed that different disturbances are independent of one-another, although such assumptions are sometimes made in particular models.[2]
- The remaining variables on the right-hand side of the model are *exogenous variables*, whose values are treated as conditionally fixed; an additional defining characteristic of exogenous variables is that they are assumed to be independent of the errors (much as the predictors in a common regression model are taken to be independent of the error).
- The γ's (Greek *gamma*) are structural parameters (regression coefficients) relating the endogenous variables to the exogenous variables (including an implicit constant regressor for each of the first three equations).
- Similarly, the β's (Greek *beta*) are structural parameters relating the endogenous variables to one-another.
- The last three equations have no error variables and no structural parameters. These equations are *identities*, and could be substituted out of the model. Our task is to estimate the first three equations, which contain unknown parameters.

The variables in model (1) have the following definitions:

C_t	Consumption (in year t)
I_t	Investment
W_t^p	Private wages
X_t	Equilibrium demand
P_t	Private profits
K_t	Capital stock
G_t	Government non-wage spending
T_t	Indirect business taxes and net exports
W_t^g	Government wages
A_t	Time trend, year $-$ 1931

The use of the subscript t for observations reflects the fact that Klein estimated the model with annual time-series data for the years 1921 through 1941.[3] Klein's data are in the data frame `Klein` in the `sem` library:

```
> library(sem)
> data(Klein)
> Klein
   year    c    p   wp    i k.lag    x  wg    g    t
1  1920 39.8 12.7 28.8  2.7 180.1 44.9 2.2  2.4  3.4
2  1921 41.9 12.4 25.5 -0.2 182.8 45.6 2.7  3.9  7.7
3  1922 45.0 16.9 29.3  1.9 182.6 50.1 2.9  3.2  3.9
4  1923 49.2 18.4 34.1  5.2 184.5 57.2 2.9  2.8  4.7
```

[1] Some forms of structural equation models do not require that one endogenous variable in each equation be indentified as the response variable.

[2] See, for example, the discussion of recursive models below.

[3] Estimating a structural equation model for time-series data raises the issue of autocorrelated errors, as it does in regression models fit to time-series data (described in the Appendix on time-series regression). Although I will not address this complication, there are methods for accommodating autocorrelated errors in structural equation models; see, e.g., Greene (1993: 608–609).

```
5  1924 50.6 19.4 33.9  3.0 189.7 57.1 3.1  3.5  3.8
6  1925 52.6 20.1 35.4  5.1 192.7 61.0 3.2  3.3  5.5
7  1926 55.1 19.6 37.4  5.6 197.8 64.0 3.3  3.3  7.0
8  1927 56.2 19.8 37.9  4.2 203.4 64.4 3.6  4.0  6.7
9  1928 57.3 21.1 39.2  3.0 207.6 64.5 3.7  4.2  4.2
10 1929 57.8 21.7 41.3  5.1 210.6 67.0 4.0  4.1  4.0
11 1930 55.0 15.6 37.9  1.0 215.7 61.2 4.2  5.2  7.7
12 1931 50.9 11.4 34.5 -3.4 216.7 53.4 4.8  5.9  7.5
13 1932 45.6  7.0 29.0 -6.2 213.3 44.3 5.3  4.9  8.3
14 1933 46.5 11.2 28.5 -5.1 207.1 45.1 5.6  3.7  5.4
15 1934 48.7 12.3 30.6 -3.0 202.0 49.7 6.0  4.0  6.8
16 1935 51.3 14.0 33.2 -1.3 199.0 54.4 6.1  4.4  7.2
17 1936 57.7 17.6 36.8  2.1 197.7 62.7 7.4  2.9  8.3
18 1937 58.7 17.3 41.0  2.0 199.8 65.0 6.7  4.3  6.7
19 1938 57.5 15.3 38.2 -1.9 201.8 60.9 7.7  5.3  7.4
20 1939 61.6 19.0 41.6  1.3 199.9 69.5 7.8  6.6  8.9
21 1940 65.0 21.1 45.0  3.3 201.2 75.7 8.0  7.4  9.6
22 1941 69.7 23.5 53.3  4.9 204.5 88.4 8.5 13.8 11.6
```

The data in `Klein` conform to my usual practice of using lower-case names for variables. Some of the variables in Klein's model have to be constructed from the data:

```
> attach(Klein)
> p.lag <- c(NA, p[-length(p)])
> x.lag <- c(NA, x[-length(x)])
> a <- year - 1931

> cbind(year, a, p, p.lag, x, x.lag)
      year   a    p p.lag    x x.lag
 [1,] 1920 -11 12.7    NA 44.9    NA
 [2,] 1921 -10 12.4  12.7 45.6  44.9
 [3,] 1922  -9 16.9  12.4 50.1  45.6
 [4,] 1923  -8 18.4  16.9 57.2  50.1
 [5,] 1924  -7 19.4  18.4 57.1  57.2
 [6,] 1925  -6 20.1  19.4 61.0  57.1
 [7,] 1926  -5 19.6  20.1 64.0  61.0
 [8,] 1927  -4 19.8  19.6 64.4  64.0
 [9,] 1928  -3 21.1  19.8 64.5  64.4
[10,] 1929  -2 21.7  21.1 67.0  64.5
[11,] 1930  -1 15.6  21.7 61.2  67.0
[12,] 1931   0 11.4  15.6 53.4  61.2
[13,] 1932   1  7.0  11.4 44.3  53.4
[14,] 1933   2 11.2   7.0 45.1  44.3
[15,] 1934   3 12.3  11.2 49.7  45.1
[16,] 1935   4 14.0  12.3 54.4  49.7
[17,] 1936   5 17.6  14.0 62.7  54.4
[18,] 1937   6 17.3  17.6 65.0  62.7
[19,] 1938   7 15.3  17.3 60.9  65.0
[20,] 1939   8 19.0  15.3 69.5  60.9
[21,] 1940   9 21.1  19.0 75.7  69.5
[22,] 1941  10 23.5  21.1 88.4  75.7
```

Notice, in particular how the lagged variables P_{t-1} and X_{t-1} are created by shifting P_t and X_t forward one time period — placing an `NA` at the beginning of each variable, and dropping the last observation. The first observation for P_{t-1} and X_{t-1} is missing because there are no data available for P_0 and X_0.

Estimating Klein's model is complicated by the presence of endogenous variables on the right-hand side of the structural equations. In general, we cannot assume that an endogenous predictor is uncorrelated with the error variable in a structural equation, and consequently ordinary least-squares (OLS) regression cannot be relied upon to produce consistent estimates of the parameters of the equation. For example, the endogenous variable P_t appears as a predictor in the first structural equation, for C_t; but X_t is a component of P_t, and X_t, in turn, depends upon C_t, one of whose components is the error ζ_{1t}. Thus, indirectly, ζ_{1t} is a component of P_t, and the two are likely correlated. Similar reasoning applies to the other endogenous predictors in the model, as a consequence of the simultaneous determination of the endogenous variables.

2.2 Identification and Instrumental-Variables Estimation

Instrumental-variables estimation provides consistent estimates of the parameters of a structural equation. An *instrumental variable* (also called an *instrument*) is a variable uncorrelated with the error of a structural equation. In the present context, the exogenous variables can serve as instrumental variables, as can predetermined endogenous variables, such as P_{t-1}.

Let us write a structural equation of the model as

$$\mathbf{y} = \mathbf{X}\boldsymbol{\delta} + \boldsymbol{\zeta} \tag{2}$$

where $\mathbf{y}$ is the $n \times 1$ vector for the response variable in the equation; $\mathbf{X}$ is an $n \times p$ model matrix, containing the p endogenous and exogenous predictors for the equation, normally including a column of 1's for the constant; $\boldsymbol{\delta}$ (Greek *delta*) is the $p \times 1$ parameter vector, containing the γ's and β's for the structural equation; and $\boldsymbol{\zeta}$ is the $n \times 1$ error vector. Let the $n \times p$ matrix $\mathbf{Z}$ contain instrumental variables (again, normally including a column of 1's). Then, multiplying the structural equation through by $\mathbf{Z}'$ produces

$$\mathbf{Z}'\mathbf{y} = \mathbf{Z}'\mathbf{X}\boldsymbol{\delta} + \mathbf{Z}'\boldsymbol{\zeta}$$

In the probability limit, $\frac{1}{n}\mathbf{Z}'\boldsymbol{\zeta}$ goes to $\mathbf{0}$ because of the uncorrelation of the instrumental variables with the error. Consequently, the instrumental-variables estimator

$$\widehat{\boldsymbol{\delta}} = (\mathbf{Z}'\mathbf{X})^{-1}\mathbf{Z}'\mathbf{y}$$

is a consistent estimator of $\boldsymbol{\delta}$.

I have implicitly assumed two things here: (1) that the number of instrumental variables is equal to the number of predictors p in the structural equation; and (2) that the cross-products matrix $\mathbf{Z}'\mathbf{X}$ is nonsingular.

- If there are *fewer* instrumental variables than predictors (i.e., structural coefficients), then the estimating equations
$$\mathbf{Z}'\mathbf{y} = \mathbf{Z}'\mathbf{X}\widehat{\boldsymbol{\delta}}$$
are under-determined, and the structural equation is said to be *under-identified*.[4]

- If there are p instrumental variables, then the structural equation is said to be *just-identified*.

- If there are *more* instrumental variables than predictors, then the estimating equations will almost surely be over-determined, and the structural equation is said to be *over-identified*.[5] What we have here is an embarrassment of riches, however: We could obtain consistent estimates simply by discarding surplus instrumental variables. To do so would be statistically profligate, however, and there are better solutions to over-identification, including the method of two-stage least squares, to be described presently.

- For $\mathbf{Z}'\mathbf{X}$ to be nonsingular, the instrumental variables must be correlated with the predictors, and we must avoid perfect collinearity.

[4] That there must be at least as many instrumental variables as coefficients to estimate in a structural equation is called the *order condition for identification*. It turns out that the order condition is a necessary, but not sufficient, condition for identification. Usually, however, a structural equation model that satisfies the order condition is identified. See the references cited in the introductory section of the appendix.

[5] This over-determination is a product of sampling error, since presumably in the population the estimating equations would hold precisely and simultaneously. If the estimating equations are highly inconsistent, that casts doubt upon the specification of the model.

2.3 Two-Stage Least Squares Estimation

Two-stage least squares (*2SLS*) is so named because it can be thought of the catenation of two OLS regression:

1. In the first stage, the predictors $\mathbf{X}$ are regressed on the instrumental variables $\mathbf{Z}$, obtaining fitted values[6]
$$\widehat{\mathbf{X}} = \mathbf{Z}(\mathbf{Z}'\mathbf{Z})^{-1}\mathbf{Z}'\mathbf{X}$$

2. In the second stage, the response $\mathbf{y}$ is regressed on the fitted values from the first stage, $\widehat{\mathbf{X}}$, producing the 2SLS estimator of $\boldsymbol{\delta}$:
$$\widehat{\boldsymbol{\delta}} = (\widehat{\mathbf{X}}'\widehat{\mathbf{X}})^{-1}\widehat{\mathbf{X}}'\mathbf{y}$$
This is justified because as linear combinations of the instrumental variables, the columns of $\widehat{\mathbf{X}}$ are (in the probability limit) uncorrelated with the structural disturbances. An alternative, but equivalent, approach to the second stage is to apply the fitted values from the first stage, $\widehat{\mathbf{X}}$, as instrumental variables to the structural equation (2):[7]
$$\widehat{\boldsymbol{\delta}} = (\widehat{\mathbf{X}}'\mathbf{X})^{-1}\widehat{\mathbf{X}}'\mathbf{y}$$

The two stages of 2SLS can be combined algebraically, producing the following expression for the estimates:
$$\widehat{\boldsymbol{\delta}} = [\mathbf{X}'\mathbf{Z}(\mathbf{Z}'\mathbf{Z})^{-1}\mathbf{Z}'\mathbf{X}]^{-1}\mathbf{X}'\mathbf{Z}(\mathbf{Z}'\mathbf{Z})^{-1}\mathbf{Z}'\mathbf{y}$$
The estimated asymptotic covariance matrix of the coefficients is
$$\widehat{\mathcal{V}}(\widehat{\boldsymbol{\delta}}) = s^2[\mathbf{X}'\mathbf{Z}(\mathbf{Z}'\mathbf{Z})^{-1}\mathbf{Z}'\mathbf{X}]^{-1}$$
where s^2 is the estimated error variance for the structural equation,
$$s^2 = \frac{(\mathbf{y} - \mathbf{X}\widehat{\boldsymbol{\delta}})'(\mathbf{y} - \mathbf{X}\widehat{\boldsymbol{\delta}})}{n - p}$$
that is, the sum of squared residuals divided by residual degrees of freedom.[8]

To apply 2SLS to the structural equations in Klein's model, we may use the four exogenous variables, the constant, and the three predetermined endogenous variables as instruments. Because there are therefore eight instrumental variables and only four structural parameters to estimate in each equation, the three structural equations are all over-identified.

The `tsls` function in the `sem` library performs 2SLS estimation:

- The structural equation to be estimated is specified by a model formula, as for `lm` (see Chapter 4 of the text).
- The instrumental variables are supplied in a one-sided model formula via the `instruments` argument
- There are optional `data`, `subset`, `na.action`, and `contrasts` arguments that work just like those in `lm` (and which are, again, described in Chapter 4 of the text).
- The `tsls` function returns an object of class `"tsls"`. A variety of methods exist for objects of this class, including `print`, `summary`, `fitted`, `residuals`, and `anova` methods. For details, enter `help(tsls)`.

For example, to estimate the structural equations in Klein's model:

[6] Columns of $\mathbf{X}$ corresponding to exogenous predictors are simply reproduced in $\widehat{\mathbf{X}}$, since the exogenous variables are among the instrumental variables in $\mathbf{Z}$.

[7] Obviously, for the two approaches to be equivalent, it must be the case that $\widehat{\mathbf{X}}'\widehat{\mathbf{X}} = \widehat{\mathbf{X}}'\mathbf{X}$. Can you see why this equation holds?

[8] Because the result is asymptotic, a less conservative alternative is to divide the residual sum of squares by n rather than by $n - p$.

```
> eqn.1 <- tsls(c ~ p + p.lag + I(wp + wg),
+     instruments= ~ g + t + wg + a + p.lag + k.lag + x.lag)

> summary(eqn.1)

 2SLS Estimates

Model Formula: c ~ p + p.lag + I(wp + wg)

Instruments: ~g + t + wg + a + p.lag + k.lag + x.lag

Residuals:
     Min.   1st Qu.    Median      Mean   3rd Qu.      Max.
-1.89e+00 -6.16e-01 -2.46e-01 -4.34e-11  8.85e-01  2.00e+00

            Estimate Std. Error t value  Pr(>|t|)
(Intercept) 16.55476    1.46798 11.2772 2.587e-09
p            0.01730    0.13120  0.1319 8.966e-01
p.lag        0.21623    0.11922  1.8137 8.741e-02
I(wp + wg)   0.81018    0.04474 18.1107 1.505e-12

Residual standard error: 1.1357 on 17 degrees of freedom

> eqn.2 <- tsls(i ~ p + p.lag + k.lag,
+     instruments= ~ g + t + wg + a + p.lag + k.lag + x.lag)

> summary(eqn.2)

 2SLS Estimates

Model Formula: i ~ p + p.lag + k.lag

Instruments: ~g + t + wg + a + p.lag + k.lag + x.lag

Residuals:
     Min.   1st Qu.    Median      Mean   3rd Qu.      Max.
-3.29e+00 -8.07e-01  1.42e-01  1.36e-11  8.60e-01  1.80e+00

            Estimate Std. Error t value Pr(>|t|)
(Intercept)  20.2782    8.38325  2.4189 0.027071
p             0.1502    0.19253  0.7802 0.445980
p.lag         0.6159    0.18093  3.4044 0.003375
k.lag        -0.1578    0.04015 -3.9298 0.001080

Residual standard error: 1.3071 on 17 degrees of freedom

> eqn.3 <- tsls(wp ~ x + x.lag + a,
+     instruments= ~ g + t + wg + a + p.lag + k.lag + x.lag)

> summary(eqn.3)

 2SLS Estimates

Model Formula: wp ~ x + x.lag + a
```

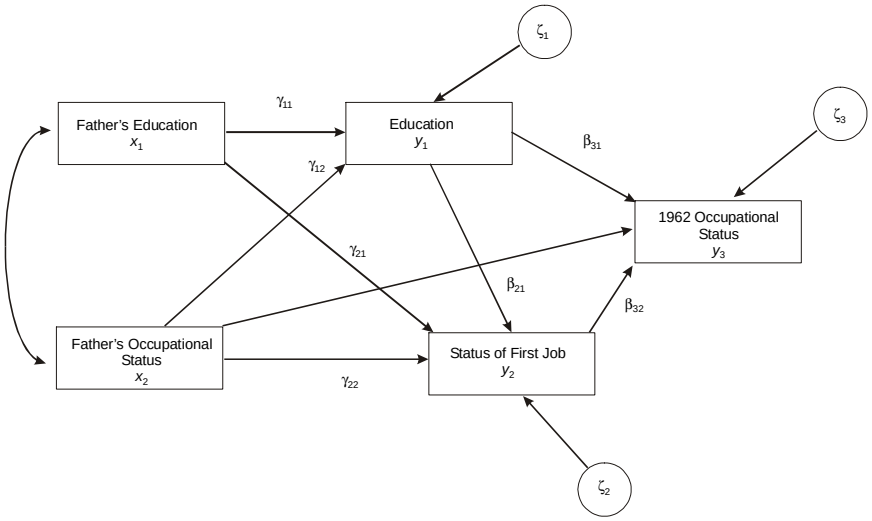

Figure 1: Blau and Duncan recursive basic stratification model.

```
Instruments: ~g + t + wg + a + p.lag + k.lag + x.lag

Residuals:
     Min.   1st Qu.    Median      Mean   3rd Qu.      Max.
-1.29e+00 -4.73e-01  1.45e-02  1.79e-11  4.49e-01  1.20e+00

            Estimate Std. Error t value  Pr(>|t|)
(Intercept)   1.5003    1.27569   1.176 2.558e-01
x             0.4389    0.03960  11.082 3.368e-09
x.lag         0.1467    0.04316   3.398 3.422e-03
a             0.1304    0.03239   4.026 8.764e-04

Residual standard error: 0.7672 on 17 degrees of freedom
```

Note the use of the identity function `I` to 'protect' the expression `wp + wg` in the first structural equation; as in a linear model, leaving an expression like this unprotected would cause the plus sign to be interpreted as specifying separate terms for the model, rather than as the sum of `wp` and `wg`, which is what is desired here.

2.4 Recursive Models

Outside of economics, it is common to specify a structural equation model in the form of a graph called a *path diagram*. A well known example, Blau and Duncan's (1967) basic stratification model, appears in Figure 1.

The following conventions, some of them familiar from Klein's macroeconomic model, are employed in drawing the path diagram:

- Directly observable variables are enclosed in rectangular boxes.
- Unobservable variables are enclosed in circles (more generally, in ellipses); in this model, the only unobservable variables are the disturbances.
- Exogenous variables are represented by x's; endogenous variables by y's; and disturbances by ζ's.

- Directed (i.e., single-headed) arrows represent structural parameters. The endogenous variables are distinguished from the exogenous variables by having directed arrows pointing towards them, while exogenous variables appear only at the tails of directed arrows.
- Bidirectional (double-headed) arrows represent non-causal, potentially nonzero, covariances between exogenous variables (and, more generally, also between disturbances).
- As before, γ's are used for structural parameters relating an endogenous to an exogenous variable, while β's are used for structural parameters relating one endogenous variable to another.
- To the extent possible, horizontal ordering of the variables corresponds to their causal ordering:[9] Thus, 'causes' appear to the left of 'effects.'

The structural equations of the model may be read off the path diagram:[10]

$$\begin{aligned} y_{1i} &= \gamma_{10} + \gamma_{11}x_{1i} + \gamma_{12}x_{2i} + \zeta_{1i} \\ y_{2i} &= \gamma_{20} + \gamma_{21}x_{1i} + \gamma_{22}x_{2i} + \beta_{21}y_{1i} + \zeta_{2i} \\ y_{3i} &= \gamma_{30} + \gamma_{32}x_{2i} + \beta_{31}y_{1i} + \beta_{32}y_{2i} + \zeta_{2i} \end{aligned}$$

Blau and Duncan's model is a member of a special class of SEMs called *recursive models*. Recursive models have the following two defining characteristics:

1. There are no reciprocal directed paths or feedback loops in the path diagram.
2. Different disturbances are independent of one-another (and hence are unlinked by bidirectional arrows).

As a consequence of these two properties, the predictors in a structural equation of a recursive model are always independent of the error of that equation, and the structural equation may be estimated by OLS regression. Estimating a recursive model is simply a sequence of OLS regressions. In S, we would of course use `lm` to fit the regressions. This is a familiar operation, and therefore I will not pursue the example further.

Structural equation models that are not recursive are sometimes termed *nonrecursive* (an awkward and often-confused adjective).

3 General Structural Equation Models

General structural equation models include unobservable exogenous or endogenous variables (also termed *factors* or *latent variables*) in addition to the unobservable disturbances. General structural equation models are sometimes called *LISREL models*, after the first widely available computer program capable of estimating this class of models (Jöreskog, 1973); LISREL is an acronym for *li*near *s*tructural *rel*ations.

Figure 2 shows the path diagram for an illustrative general structural equation model, from path-breaking work by Duncan, Haller, and Portes (1968) concerning peer influences on the aspirations of male high-school students. The most striking new feature of this model is that two of the endogenous variables, Respondent's General Aspirations (η_1) and Friend's General Aspirations (η_2), are unobserved variables. Each of these variables has two observed indicators: The occupational and educational aspirations of each boy — y_1 and y_2 for the respondent, and y_3 and y_4 for his best friend.

3.1 The LISREL Model

It is common in general structural equation models such as the peer-influences model to distinguish between two sub-models:

[9] When there are feedback loops in a model, it is impossible to satisfy the left-to-right rule without using curved directed arrows.

[10] In writing out the structural equations from a path diagram, it is common to omit the intercept parameters (here, γ_{10}, γ_{20}, and γ_{30}), for which no paths appear. To justify this practice, we may express all variables as deviations from their expectations (in the sample, as deviations from their means), eliminating the intercept from each regression equation.

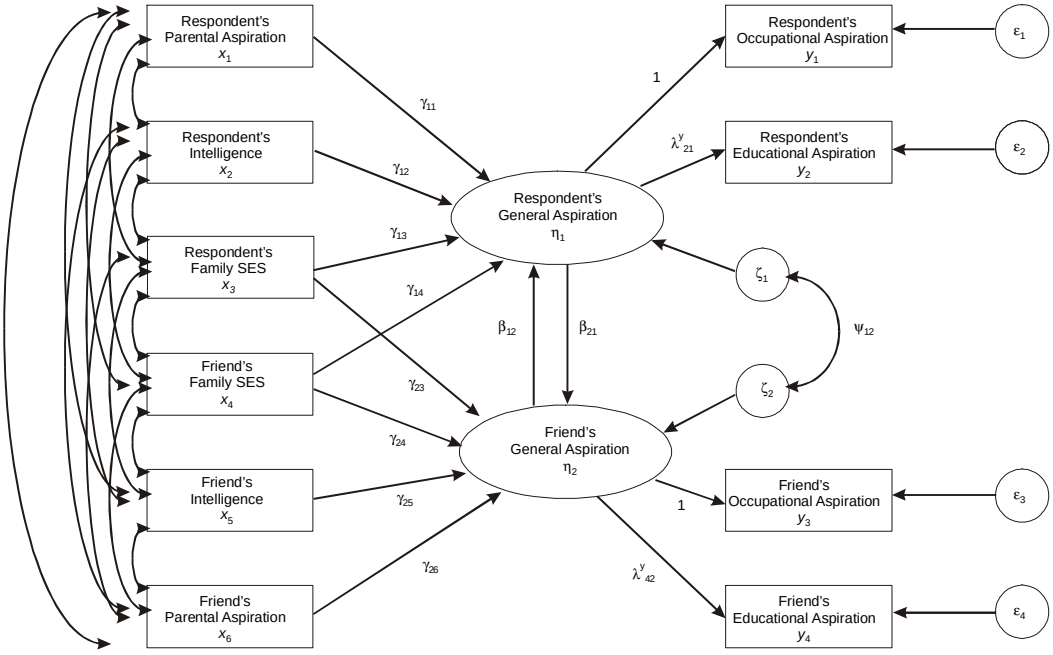

Figure 2: Duncan, Haller, and Portes's general structural equation model for peer influences on aspirations.

1. A structural submodel, relating endogenous to exogenous variables and to one-another. In the peer-influences model, the endogenous variables are unobserved, while the exogenous variables are directly observed.

2. A measurement submodel, relating latent variables (here only latent endogenous variables) to their indicators.

I have used the following notation, associated with Jöreskog's LISREL model, in drawing the path diagram in Figure 2:

- x's are used to represent observable exogenous variables. If there were *latent* exogenous variables in the model, these would be represented by ξ's (Greek *xi*), and x's would be used to represent their observable indicators.

- y's are employed to represent the indicators of the latent endogenous variables, which are symbolized by η's (Greek *eta*). Were there directly observed endogenous variables in the model, then these too would be represented by y's.

- As before, γ's and β's are used, respectively, for structural coefficients relating endogenous variables to exogenous variables and to one-another, and ζ's are used for structural disturbances. The parameter ψ_{12} is the covariance between the disturbances ζ_1 and ζ_2. The variances of the disturbances, ψ_1^2 and ψ_2^2, are not shown on the diagram.

- In the measurement submodel, λ's (Greek *lambda*) represent regression coefficients (also called *factor loadings*) relating observable indicators to latent variables. The superscript y in λ^y indicates that the factor loadings in this model pertain to indicators of latent *endogenous* variables. Notice that one λ for each factor is set to 1; this is done to identify the scale of the corresponding latent variable.

- The ε's (Greek *epsilon*) represent measurement error in the endogenous indicators; if there were *exogenous* indicators in the model, then the measurement errors associated with them would be represented by δ's (Greek *delta*).

Symbol	Meaning
N	Number of observations
m	Number of latent endogenous variables
n	Number of latent exogenous variables
p	Number of indicators of latent endogenous variables
q	Number of indicators of latent exogenous variable
$\underset{(m\times 1)}{\boldsymbol{\eta}_i}$	Latent endogenous variables (for observation i)
$\underset{(n\times 1)}{\boldsymbol{\xi}_i}$	Latent exogenous variables
$\underset{(m\times 1)}{\boldsymbol{\varsigma}_i}$	Structural disturbances (errors in equations)
$\underset{(m\times m)}{\mathbf{B}}$	Structural parameters relating latent endogenous variables
$\underset{(m\times n)}{\boldsymbol{\Gamma}}$	Structural parameters relating latent endogenous to exogenous variables
$\underset{(p\times 1)}{\mathbf{y}_i}$	Indicators of latent endogenous variables
$\underset{(q\times 1)}{\mathbf{x}_i}$	Indicators of latent exogenous variables
$\underset{(p\times 1)}{\boldsymbol{\varepsilon}_i}$	Measurement errors in endogenous indicators
$\underset{(q\times 1)}{\boldsymbol{\delta}_i}$	Measurement errors in exogenous indicators
$\underset{(p\times m)}{\boldsymbol{\Lambda}_y}$, $\underset{(q\times n)}{\boldsymbol{\Lambda}_x}$	Factor loadings relating indicators to latent variables
$\underset{(n\times n)}{\Phi}$	Covariances among latent exogenous variables
$\underset{(m\times m)}{\Psi}$	Covariances among structural disturbances
$\underset{(p\times p)}{\boldsymbol{\Theta}_\varepsilon}$, $\underset{(q\times q)}{\boldsymbol{\Theta}_\delta}$	Covariances among measurement errors
$\underset{(p+q\times p+q)}{\boldsymbol{\Sigma}}$	Covariances among observed (indicator) variables

Table 1: Notation for the LISREL model. The order of each vector or matrix is shown in paretheses below its symbol.

We are swimming in notation, but we still require some more (not all of which is necessary for the peer-influences model): We use σ_{ij} (Greek *sigma*) to represent the covariance between two observable variables; $\theta^{\varepsilon}_{ij}$ to represent the covariance between two measurement-error variables for endogenous indicators, ε_i and ε_j; θ^{δ}_{ij} to represent the covariance between two measurement-error variables for exogenous indicators, δ_i and δ_j; and ϕ_{ij} to represent the covariance between two latent exogenous variables ξ_i and ξ_j.

The LISREL notation for general structural equation models is summarized in Table 1. The structural and measurement submodels are written as follows:

$$\begin{aligned}
\boldsymbol{\eta}_i &= \mathbf{B}\boldsymbol{\eta}_i + \boldsymbol{\Gamma}\boldsymbol{\xi}_i + \boldsymbol{\varsigma}_i \\
\mathbf{y}_i &= \boldsymbol{\Lambda}_y\boldsymbol{\eta}_i + \boldsymbol{\varepsilon}_i \\
\mathbf{x}_i &= \boldsymbol{\Lambda}_x\boldsymbol{\xi}_i + \boldsymbol{\delta}_i
\end{aligned}$$

In order to identify the model, many of the parameters in $\mathbf{B}, \boldsymbol{\Gamma}, \boldsymbol{\Lambda}_x, \boldsymbol{\Lambda}_y, \boldsymbol{\Phi}, \boldsymbol{\Psi}, \boldsymbol{\Theta}_\varepsilon$, and $\boldsymbol{\Theta}_\delta$ must be constrained, typically by setting parameters to 0 or 1, or by defining certain parameters to be equal.

3.2 The RAM Formulation

Although LISREL notation is most common, there are several equivalent ways to represent general structural equation models. The `sem` function uses the simpler *RAM* (*reticular action model* – don't ask!) formulation of McArdle (1980) and McArdle and McDonald (1984); the notation that I employ below is from McDonald and Hartmann (1992).

The RAM model includes two vectors of variables: $\mathbf{v}$, which contains the indicator variables, directly observed exogenous variables, and the latent exogenous and endogenous variables in the model; and $\mathbf{u}$, which contains directly observed exogenous variables, measurement-error variables, and structural disturbances. The two sets of variables are related by the equation

$$\mathbf{v} = \mathbf{A}\mathbf{v} + \mathbf{u}$$

Thus, the matrix $\mathbf{A}$ includes structural coefficients and factor loadings. For example, for the Duncan, Haller, and Portes model, we have (using LISREL notation for the individual parameters):

$$
\begin{bmatrix} x_1 \\ x_2 \\ x_3 \\ x_4 \\ x_5 \\ x_6 \\ y_1 \\ y_2 \\ y_3 \\ y_4 \\ \eta_1 \\ \eta_2 \end{bmatrix} =
\begin{bmatrix}
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & \lambda^y_{21} & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & \lambda^y_{42} \\
\gamma_{11} & \gamma_{12} & \gamma_{13} & \gamma_{14} & 0 & 0 & 0 & 0 & 0 & 0 & 0 & \beta_{12} \\
0 & 0 & \gamma_{23} & \gamma_{24} & \gamma_{25} & \gamma_{26} & 0 & 0 & 0 & 0 & \beta_{21} & 0
\end{bmatrix}
\begin{bmatrix} x_1 \\ x_2 \\ x_3 \\ x_4 \\ x_5 \\ x_6 \\ y_1 \\ y_2 \\ y_3 \\ y_4 \\ \eta_1 \\ \eta_2 \end{bmatrix} +
\begin{bmatrix} x_1 \\ x_2 \\ x_3 \\ x_4 \\ x_5 \\ x_6 \\ \varepsilon_1 \\ \varepsilon_2 \\ \varepsilon_3 \\ \varepsilon_4 \\ \zeta_1 \\ \zeta_2 \end{bmatrix}
$$

It is typically the case that $\mathbf{A}$ is sparse, containing many 0's. Notice the special treatment of the observed exogenous variables, x_1 through x_6, which are specified to be measured without error, and which consequently appear both in $\mathbf{v}$ and $\mathbf{u}$.

The final component of the RAM formulation is the covariance matrix $\mathbf{P}$ of $\mathbf{u}$. Assuming that all of the error variables have expectations of 0, and that all other variables have been expressed as deviations from their expectations, $\mathbf{P} = E(\mathbf{uu'})$. For the illustrative model,

$$
\mathbf{P} = \begin{bmatrix}
\sigma_{11} & \sigma_{12} & \sigma_{13} & \sigma_{14} & \sigma_{15} & \sigma_{16} & 0 & 0 & 0 & 0 & 0 & 0 \\
\sigma_{21} & \sigma_{22} & \sigma_{23} & \sigma_{24} & \sigma_{25} & \sigma_{26} & 0 & 0 & 0 & 0 & 0 & 0 \\
\sigma_{31} & \sigma_{32} & \sigma_{33} & \sigma_{34} & \sigma_{35} & \sigma_{36} & 0 & 0 & 0 & 0 & 0 & 0 \\
\sigma_{41} & \sigma_{42} & \sigma_{43} & \sigma_{44} & \sigma_{45} & \sigma_{46} & 0 & 0 & 0 & 0 & 0 & 0 \\
\sigma_{51} & \sigma_{52} & \sigma_{53} & \sigma_{54} & \sigma_{55} & \sigma_{56} & 0 & 0 & 0 & 0 & 0 & 0 \\
\sigma_{61} & \sigma_{62} & \sigma_{63} & \sigma_{64} & \sigma_{65} & \sigma_{66} & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & \theta^\varepsilon_{11} & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & \theta^\varepsilon_{22} & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & \theta^\varepsilon_{33} & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & \theta^\varepsilon_{44} & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & \psi_{11} & \psi_{12} \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & \psi_{21} & \psi_{22}
\end{bmatrix}
$$

For convenience, I use a double-subscript notation for both covariances and variances; thus, for example, σ_{11} is the variance of x_1 (usually written σ^2_1); θ^ε_{11} is the variance of ε_1; and ψ_{11} is the variance of ζ_1.

The key to estimating the model is the connection between the covariances of the observed variables, which may be estimated directly from sample data, and the parameters in $\mathbf{A}$ and $\mathbf{P}$. Let m denote the the number of variables in $\mathbf{v}$, and (without loss of generality) let the first n of these be the observed variables in the model.[11] Define the $m \times m$ *selection matrix* $\mathbf{J}$ to pick out the observed variables; that is

$$
\mathbf{J} = \begin{bmatrix} \mathbf{I}_n & \mathbf{0} \\ \mathbf{0} & \mathbf{0} \end{bmatrix}
$$

where $\mathbf{I}_n$ is the order$-n$ identity matrix, and the $\mathbf{0}$'s are zero matrices of appropriate orders. The model implies the following covariances among the observed variables:

$$
\mathbf{C} = E(\mathbf{Jvv'J'}) = \mathbf{J}(\mathbf{I}_m - \mathbf{A})^{-1}\mathbf{P}(\mathbf{I}_m - \mathbf{A})^{-1'}\mathbf{J}'
$$

Let $\mathbf{S}$ denote the observed-variable covariances computed directly from the sample. Fitting the model to the data — that is, estimating the free parameters in $\mathbf{A}$ and $\mathbf{P}$ — entails selecting parameter values that make $\mathbf{S}$ as close as possible to the model-implied covariances $\mathbf{C}$. Under the assumptions that the errors and latent variables are multivariately normally distributed, finding the maximum-likelihood estimates of the free parameters in $\mathbf{A}$ and $\mathbf{P}$ is equivalent to minimizing the criterion

$$
F(\mathbf{A}, \mathbf{P}) = \text{trace}(\mathbf{SC}^{-1}) - n + \log_e \det \mathbf{C} - \log_e \det \mathbf{S} \tag{3}
$$

[11] Notice the nonstandard use of n to represent the number of observed variables rather than the sample size. The latter is represented by N, as in the LISREL model.

3.3 The sem Function

The sem function computes maximum-likelihood estimates for general structural equation models, using the RAM formulation of the model. There are three required arguments to sem:

1. ram: a specification of the single and double-headed arrows in the model, corresponding to elements in the parameter matrices **A** and **P**.

   (a) The first column of ram specifies parameters in the form "A -> B" for a single-headed arrow from variable A to variable B, and "A <-> B" for the covariance between variables A and B; a variance is specified as "A <-> A". If a variable name (e.g., A) is not among the observed variables in the covariance matrix (see the second argument to sem, immediately below), then it is assumed to represent a latent variable.[12] Error variables do not appear explicitly, but error variances and covariances are specified using the names of the corresponding endogenous variables; for example, "B <-> B" is the variance of the disturbance associated with B, if B is an endogenous variable, or the measurement-error variance of B, if B is an indicator.[13]

   (b) The second column of ram supplies a name for the parameter corresponding to the arrow — for example, "beta12". If two or more arrows are given the same name, then their parameters are constrained to be equal. If a parameter is to be fixed at a non-zero value rather than estimated, then its name is NA.

   (c) The last column of ram gives a start-value for the parameter. If this value is NA, then the program will calculate a start-value using an adaptation of the method described by McDonald and Hartmann (1992). If the parameter is fixed, then a value *must* be supplied. At present, the method used to calculate start-values is not reliable, and may cause sem to fail, reporting either a singular matrix or a non-finite objective function. In cases like this, you can request that the start-values and some other diagnostic output be printed, by including the argument debug=T; after examining the diagnostic output, you might try to specify some of the start-values directly.

   (d) If there are fixed exogenous variables in the model (such as variables x_1 through x_6 in the peer-influences model), then the variances and covariances of these variables do not have to be specified explicitly in the ram argument to sem. Rather, the names of the fixed exogenous variables can be supplied via the argument fixed.x.

2. S: the sample covariance matrix among the observed variables in the model. The covariances may be obtained from a secondary source or computed by the standard S function var. If S has row and column names, then these are used by default as the names of the observed variables. The sem function accepts a lower or upper-triangular covariance matrix, as well as the full (symmetric) covariance matrix.

3. N: the sample size on which the covariance matrix **S** is based.

Enter help(sem) for a description of the optional arguments to sem.

The Duncan, Haller and Portes model was estimated for standardized variables, so the input covariance matrix is a correlation matrix:[14]

```
> R.dhp <- matrix(c(      # lower triangle of correlation matrix
+     1,      0,      0,      0,      0,      0,      0,      0,      0,      0,
+     .6247,   1,     0,      0,      0,      0,      0,      0,      0,      0,
+     .3269,  .3669,  1,      0,      0,      0,      0,      0,      0,      0,
+     .4216,  .3275,  .6404,  1,      0,      0,      0,      0,      0,      0,
+     .2137,  .2742,  .1124,  .0839,  1,      0,      0,      0,      0,      0,
```

[12] A consequence of this convention is that typing errors inadvertently create latent variables.

[13] Indeed, it is not necessary in the RAM formulation to draw a distinction between indicators and (other) endogenous variables.

[14] Using correlation-matrix input raises a complication: The standard deviations employed to standardize variables are estimated from the data, and are therefore an additional source of uncertainty in the estimates of the standardized coefficients. I will simply bypass this issue, however, which is tantamount to analyzing the data on scales conditional on the sample standard deviations.

```
+     .4105,  .4043,  .2903,  .2598,  .1839,  1,      0,      0,      0,      0,
+     .3240,  .4047,  .3054,  .2786,  .0489,  .2220,  1,      0,      0,      0,
+     .2930,  .2407,  .4105,  .3607,  .0186,  .1861,  .2707,  1,      0,      0,
+     .2995,  .2863,  .5191,  .5007,  .0782,  .3355,  .2302,  .2950,  1,      0,
+     .0760,  .0702,  .2784,  .1988,  .1147,  .1021,  .0931, -.0438,  .2087,  1
+     ), ncol=10, byrow=T)

> rownames(R.dhp) <- colnames(R.dhp) <- c('ROccAsp', 'REdAsp', 'FOccAsp',
+     'FEdAsp', 'RParAsp', 'RIQ', 'RSES', 'FSES', 'FIQ', 'FParAsp')

> R.dhp
        ROccAsp REdAsp FOccAsp FEdAsp RParAsp    RIQ   RSES    FSES    FIQ FParAsp
ROccAsp  1.0000 0.0000  0.0000 0.0000  0.0000 0.0000 0.0000  0.0000 0.0000       0
REdAsp   0.6247 1.0000  0.0000 0.0000  0.0000 0.0000 0.0000  0.0000 0.0000       0
FOccAsp  0.3269 0.3669  1.0000 0.0000  0.0000 0.0000 0.0000  0.0000 0.0000       0
FEdAsp   0.4216 0.3275  0.6404 1.0000  0.0000 0.0000 0.0000  0.0000 0.0000       0
RParAsp  0.2137 0.2742  0.1124 0.0839  1.0000 0.0000 0.0000  0.0000 0.0000       0
RIQ      0.4105 0.4043  0.2903 0.2598  0.1839 1.0000 0.0000  0.0000 0.0000       0
RSES     0.3240 0.4047  0.3054 0.2786  0.0489 0.2220 1.0000  0.0000 0.0000       0
FSES     0.2930 0.2407  0.4105 0.3607  0.0186 0.1861 0.2707  1.0000 0.0000       0
FIQ      0.2995 0.2863  0.5191 0.5007  0.0782 0.3355 0.2302  0.2950 1.0000       0
FParAsp  0.0760 0.0702  0.2784 0.1988  0.1147 0.1021 0.0931 -0.0438 0.2087       1
```

The `ram` specification may be read off the path diagram (Figure 2), remembering that the error variables do not appear explicitly, and that we do not have to supply variances and covariances for the six fixed exogenous variables:

```
> ram.dhp <- matrix(c(
+ #   arrow                       parameter       start-value
+     'RParAsp  -> RGenAsp',      'gam11',        NA,
+     'RIQ      -> RGenAsp',      'gam12',        NA,
+     'RSES     -> RGenAsp',      'gam13',        NA,
+     'FSES     -> RGenAsp',      'gam14',        NA,
+     'RSES     -> FGenAsp',      'gam23',        NA,
+     'FSES     -> FGenAsp',      'gam24',        NA,
+     'FIQ      -> FGenAsp',      'gam25',        NA,
+     'FParAsp  -> FGenAsp',      'gam26',        NA,
+     'FGenAsp  -> RGenAsp',      'bet12',        NA,
+     'RGenAsp  -> FGenAsp',      'bet21',        NA,
+     'RGenAsp  -> ROccAsp',      NA,             1,
+     'RGenAsp  -> REdAsp',       'lamy21',       NA,
+     'FGenAsp  -> FOccAsp',      NA,             1,
+     'FGenAsp  -> FEdAsp',       'lamy42',       NA,
+     'RGenAsp <-> RGenAsp',      'psi11',        NA,
+     'FGenAsp <-> FGenAsp',      'psi22',        NA,
+     'RGenAsp <-> FGenAsp',      'psi12',        NA,
+     'ROccAsp <-> ROccAsp',      'theps1',       NA,
+     'REdAsp  <-> REdAsp',       'theps2',       NA,
+     'FOccAsp <-> FOccAsp',      'theps3',       NA,
+     'FEdAsp  <-> FEdAsp',       'theps4',       NA),
+     ncol=3, byrow=T)
>
```

To fit the model, I note that the Duncan, Haller, and Portes data set comprises $N = 329$ observations:

```
> sem.dhp <- sem(ram.dhp, R.dhp, N=329,
+     fixed.x=c('RParAsp', 'RIQ', 'RSES', 'FSES', 'FIQ', 'FParAsp'))

> sem.dhp

 Model Chisquare =  26.69722   Df =  15

      gam11       gam12       gam13       gam14       gam23
 0.16122390  0.24965251  0.21840357  0.07184300  0.06189390
      gam24       gam25       gam26       bet12       bet21
 0.22886776  0.34903879  0.15953516  0.18422617  0.23545788
     lamy21      lamy42       psi11       psi22       psi12
 1.06267364  0.92972672  0.28098743  0.26383649 -0.02260149
     theps1      theps2      theps3      theps4
 0.41214471  0.33614760  0.31119372  0.40460356

 Iterations =  28
```

The `sem` function returns an object of class `"sem"`; the `print` method for `sem` objects displays parameter estimates, together with the likelihood-ratio chi-square statistic for the model, contrasting the model with a just-identified (or *saturated*) model, which perfectly reproduces the sample covariance matrix. The degrees of freedom for this test are equal to the degree of over-identification of the model — the difference between the number of covariances among observed variables, $n(n+1)/2$, and the number of independent parameters in the model.[15]

More information is provided by the `summary` method for `sem` objects:

```
> summary(sem.dhp)

 Model Chisquare =  26.697   Df =  15 Pr(>Chisq) = 0.031302
 Goodness-of-fit index =  0.98439
 Adjusted goodness-of-fit index =  0.94275
 RMSEA index =  0.048759   90% CI: (0.014516, 0.078314)
 BIC =  -94.782

 Normalized Residuals
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
-0.8010 -0.1180  0.0000 -0.0120  0.0398  1.5700

 Parameter Estimates
       Estimate Std Error z value   Pr(>|z|)
gam11   0.161224  0.038487  4.1890 2.8019e-05 RGenAsp <--- RParAsp
gam12   0.249653  0.044580  5.6001 2.1428e-08     RGenAsp <--- RIQ
gam13   0.218404  0.043476  5.0235 5.0730e-07    RGenAsp <--- RSES
gam14   0.071843  0.050335  1.4273 1.5350e-01    RGenAsp <--- FSES
gam23   0.061894  0.051738  1.1963 2.3158e-01    FGenAsp <--- RSES
gam24   0.228868  0.044495  5.1437 2.6938e-07    FGenAsp <--- FSES
gam25   0.349039  0.044551  7.8346 4.6629e-15     FGenAsp <--- FIQ
gam26   0.159535  0.040129  3.9755 7.0224e-05 FGenAsp <--- FParAsp
bet12   0.184226  0.096207  1.9149 5.5506e-02 RGenAsp <--- FGenAsp
bet21   0.235458  0.119742  1.9664 4.9256e-02 FGenAsp <--- RGenAsp
lamy21  1.062674  0.091967 11.5549 0.0000e+00  REdAsp <--- RGenAsp
lamy42  0.929727  0.071152 13.0668 0.0000e+00  FEdAsp <--- FGenAsp
```

[15] For the model to be identified the degrees of freedom must be 0 or greater — that is, there must be at least as many observable covariances as free parameters of the model. Unlike the order condition for the identification of observed-variable SEMs, however, it is common for this requirement to be met and yet for the model to be under-identified.

```
psi11   0.280987  0.046311  6.0674 1.2999e-09 RGenAsp <--> RGenAsp
psi22   0.263836  0.044902  5.8759 4.2067e-09 FGenAsp <--> FGenAsp
psi12  -0.022601  0.051649 -0.4376 6.6168e-01 FGenAsp <--> RGenAsp
theps1  0.412145  0.052211  7.8939 2.8866e-15 ROccAsp <--> ROccAsp
theps2  0.336148  0.053323  6.3040 2.9003e-10   REdAsp <--> REdAsp
theps3  0.311194  0.046665  6.6687 2.5800e-11 FOccAsp <--> FOccAsp
theps4  0.404604  0.046733  8.6578 0.0000e+00   FEdAsp <--> FEdAsp

 Iterations =  28
```

- The marginally significant chi-square statistic indicates that the model can be rejected. Because any over-identified model can be rejected in a sufficiently large sample, structural-equation modelers typically attend to the descriptive adequacy of the model as well as to this formal *over-identification test*.

- The *goodness-of-fit index* (*GFI*) and the *adjusted goodness-of-fit index* (*AGFI*) are ad-hoc measures of the descriptive adequacy of the model, included in the output of the `summary` method for `sem` objects because they are in common use. The GFI and AGFI are defined as follows:

$$\begin{aligned} \text{GFI} &= 1 - \frac{\text{trace}\{[\mathbf{C}^{-1}(\mathbf{S}-\mathbf{C})]^2\}}{\text{trace}[(\mathbf{C}^{-1}\mathbf{S})^2]} \\ \text{AGFI} &= 1 - \frac{n(n+1)}{2 \times \text{df}}(1 - \text{GFI}) \end{aligned}$$

  where df is the degrees of freedom for the model. Although the GFI and AGFI are thought of as proportions, comparing the value of the fitting criterion for the model with the value of the fitting criterion when no model is fit to the data, these indices are not constrained to the interval 0 to 1. Several rough cutoffs for the GFI and AGFI have been proposed; a general theme is that they should be close to 1. It is probably fair to say that the GFI and AGFI are of little pratical value.

- There is a veritable cottage industry in ad-hoc fit indices and their evaluation. See, for example, the papers in the volume edited by Bollen and Long (1993). One index that is perhaps more attractive than the others is the RMSEA (*root mean-squared error approximation*), which is an estimate of fit of the model relative to a saturated model in the population, and is computed as

$$\text{RMSEA} = \sqrt{\max\left(\frac{F}{\text{df}} - \frac{1}{N-1}, 0\right)}$$

  Here, F is the minimized fitting criterion, from equation (3). Small values of the RMSEA indicate that the model fits nearly as well as a saturated model; RMSEA ≤ 0.05 is generally taken as a good fit to the data. It is possible, moreover, to compute a confidence interval for the RMSEA. Note that the RMSEA for the peer-influences model is a bit smaller than 0.05.

- In contrast with ad-hoc fit indices, the *Bayesian information criterion* (*BIC*) has a sound statistical basis (see Raftery, 1993). The BIC adjusts the likelihood-ratio chi-square statistic L^2 for the number of parameters in the model, the number of observed variables, and the sample size:

$$\text{BIC} = L^2 - \text{df} \times \log_e nN$$

  Negative values of BIC indicate a model that has greater support from the data than the just-identified model, for which BIC is 0. *Differences* in BIC may be used to compare alternative over-identified models; indeed, the BIC is used in a variety of contexts for model selection, not just in structural-equation modeling. Raftery suggests that a BIC difference of 5 is indicative of "strong evidence" that one model is superior to another, while a difference of 10 is indicative of "conclusive evidence."

- The `sem` library provides several methods for calculating *residual covariances*, which compare the observed and model-implied covariance matrices, $\mathbf{S}$ and $\mathbf{C}$: Enter `help(residuals.sem)` for details.

The `summary` method for `sem` objects prints summary statistics for the distribution of the *normalized residual covariances*, which are defined as

$$\frac{s_{ij} - c_{ij}}{\sqrt{\frac{c_{ii}c_{jj} + c_{ij}^2}{N}}}$$

All of the structural coefficients in the peer-influences model are statistically significant, except for the coefficients linking each boy's general aspiration to the other boy's family socioeconomic status (SES).[16]

To illustrate setting parameter-equality constraints, I take advantage of the symmetry of the model to specify that all coefficients and error covariances in the top half of the path diagram (Figure 2) are the same as the corresponding parameters in the lower half.[17] These constraints are plausible in light of the parameter estimates in the initial model, since corresponding estimates have similar values. The equality constraints are imposed as follows:

```
> ram.dhp.1 <- matrix(c(
+     'RParAsp  -> RGenAsp',      'gam1',        NA,
+     'RIQ      -> RGenAsp',      'gam2',        NA,
+     'RSES     -> RGenAsp',      'gam3',        NA,
+     'FSES     -> RGenAsp',      'gam4',        NA,
+     'RSES     -> FGenAsp',      'gam4',        NA,
+     'FSES     -> FGenAsp',      'gam3',        NA,
+     'FIQ      -> FGenAsp',      'gam2',        NA,
+     'FParAsp  -> FGenAsp',      'gam1',        NA,
+     'FGenAsp  -> RGenAsp',      'bet',         NA,
+     'RGenAsp  -> FGenAsp',      'bet',         NA,
+     'RGenAsp  -> ROccAsp',      NA,             1,
+     'RGenAsp  -> REdAsp',       'lamy',        NA,
+     'FGenAsp  -> FOccAsp',      NA,             1,
+     'FGenAsp  -> FEdAsp',       'lamy',        NA,
+     'RGenAsp <-> RGenAsp',      'psi',         NA,
+     'FGenAsp <-> FGenAsp',      'psi',         NA,
+     'RGenAsp <-> FGenAsp',      'psi12',       NA,
+     'ROccAsp <-> ROccAsp',      'theps1',      NA,
+     'REdAsp  <-> REdAsp',       'theps2',      NA,
+     'FOccAsp <-> FOccAsp',      'theps1',      NA,
+     'FEdAsp  <-> FEdAsp',       'theps2',      NA),
+     ncol=3, byrow=T)

> sem.dhp.1 <- sem(ram.dhp.1, R.dhp, N=329,
+     fixed.x=c('RParAsp', 'RIQ', 'RSES', 'FSES', 'FIQ', 'FParAsp'))

> summary(sem.dhp.1)

 Model Chisquare =  32.647   Df =  24 Pr(>Chisq) = 0.11175
 Goodness-of-fit index =  0.98046
 Adjusted goodness-of-fit index =  0.95522
 RMSEA index =  0.033143   90% CI: (0, 0.059373)
 BIC =  -161.72

 Normalized Residuals
```

[16] The path from friend's to respondent's general aspiration is statistically significant by a one-sided test, which is appropriate here since the coefficient was expected to be positive.

[17] Although this specification makes some sense, the data are not entirely symmetric: Boys nominated their best friends, but this selection was not necessarily reciprocated.

```
     Min.   1st Qu.    Median      Mean   3rd Qu.      Max.
-8.78e-01 -2.05e-01 -5.00e-16 -1.67e-02  1.11e-01  1.04e+00

 Parameter Estimates
       Estimate Std Error  z value   Pr(>|z|)
gam1    0.157091  0.028245  5.56177 2.6706e-08 RGenAsp <--- RParAsp
gam2    0.301742  0.032209  9.36812 0.0000e+00     RGenAsp <--- RIQ
gam3    0.221045  0.031698  6.97343 3.0931e-12    RGenAsp <--- RSES
gam4    0.072805  0.036474  1.99609 4.5924e-02    RGenAsp <--- FSES
bet     0.204964  0.076903  2.66523 7.6935e-03 RGenAsp <--- FGenAsp
lamy    0.988764  0.054581 18.11561 0.0000e+00  REdAsp <--- RGenAsp
psi     0.274828  0.033013  8.32493 0.0000e+00 RGenAsp <--> RGenAsp
psi12  -0.014079  0.051365 -0.27410 7.8400e-01 FGenAsp <--> RGenAsp
theps1  0.360262  0.033593 10.72421 0.0000e+00 ROccAsp <--> ROccAsp
theps2  0.374557  0.033586 11.15227 0.0000e+00   REdAsp <--> REdAsp

 Iterations =  24
```

Because pairs of parameters are constrained to be equal, this model has fewer free parameters, and correspondingly more degrees of freedom, than the original model. We can perform a likelihood-ratio test for the parameter constraints by taking differences in the model chi-square statistics and degrees of freedom:

$$\begin{aligned} L^2 &= 32.647 - 26.697 = 5.950 \\ df &= 24 - 15 = 9 \\ p &= .74 \end{aligned}$$

Thus, the data appear to be consistent with the parameter constraints. Moreover, the more parsimonious constrained model has a much smaller BIC than the original model, and the constrained model has a non-significant over-identification test; the RMSEA has also improved.

So-called *modification indices* are test statistics for fixed and constrained parameters in a structural equation model. If, for example, a parameter is incorrectly constrained to 0, then the test statistic for this parameter should be large. Most commonly, modification indices are score statistics. The version implemented in the `mod.indices` function in the `sem` library is based on the likelihood-ratio statistic, refitting the model freeing each parameter in turn (but fixing all other parameters to their current values).

Applying `mod.indices` to the respecified peer-influences model produces the following result:

```
> mod.indices(sem.dhp.1)

 Approximations based on at most 10 iterations

 5 largest modification indices, A matrix:
 ROccAsp:FEdAsp  FEdAsp:ROccAsp    ROccAsp:RSES FOccAsp:ROccAsp
         3.7681          3.6722          2.7206          2.6006
FOccAsp:FParAsp
         2.2302

  5 largest modification indices, P matrix:
 FEdAsp:ROccAsp FOccAsp:ROccAsp   FEdAsp:REdAsp     RSES:REdAsp
        10.2407          8.7318          3.3861          3.3800
   RSES:ROccAsp
         3.2438
```

The `mod.indices` function returns an object of class `"sem.modind"`; the `print` method for objects of this class reports the largest modification indices for parameters in the **A** and **P** matrices of the RAM model. These are chi-square statistics, each on one degree of freedom. Because they are based on one-dimensional

optimizations, the modification indices tend to understate the improvement in the fit of the model: If the model were refit to the data, the values of other free parameters could change as well. There is also a problem of simultaneous inference in examining the largest of many test statistics. Nevertheless, the modification indices can suggest improvements to an ill-fitting model. The summary method for `sem.modind` objects prints the full matrices of modification indices, along with estimated changes in the parameter estimates upon freeing individual parameters.

Although the respecified peer-influences model fits quite well, I pursue the modification indices for the purpose of illustration. None of the modification indices for coefficients in $\mathbf{A}$ is very large, but there are a couple of moderately large modification indices for the covariances in $\mathbf{P}$. Both of these involve measurement-error covariances between indicators of general aspirations for the respondent and for the best friend. Correlated measurement errors between friend's educational aspiration and respondent's occupational aspiration (the covariance with the largest modification index) does not seem substantively compelling, but correlated errors between the two indicators of occupational aspirations (corresponding to the second-largest modification index) makes more sense.

Respecifying the model to accomodate the error correlation for the two educational-aspiration indicators yields a substantial decrease in the chi-square statistic for the model (about 10 — as expected, slightly larger than the modification index), a small decrease in the BIC, and an RMSEA of 0:

```
> ram.dhp.2 <- matrix(c(
+     'RParAsp  -> RGenAsp',      'gam1',        NA,
+     'RIQ      -> RGenAsp',      'gam2',        NA,
+     'RSES     -> RGenAsp',      'gam3',        NA,
+     'FSES     -> RGenAsp',      'gam4',        NA,
+     'RSES     -> FGenAsp',      'gam4',        NA,
+     'FSES     -> FGenAsp',      'gam3',        NA,
+     'FIQ      -> FGenAsp',      'gam2',        NA,
+     'FParAsp  -> FGenAsp',      'gam1',        NA,
+     'FGenAsp  -> RGenAsp',      'bet',         NA,
+     'RGenAsp  -> FGenAsp',      'bet',         NA,
+     'RGenAsp  -> ROccAsp',      NA,             1,
+     'RGenAsp  -> REdAsp',       'lamy',        NA,
+     'FGenAsp  -> FOccAsp',      NA,             1,
+     'FGenAsp  -> FEdAsp',       'lamy',        NA,
+     'RGenAsp <-> RGenAsp',      'psi',         NA,
+     'FGenAsp <-> FGenAsp',      'psi',         NA,
+     'RGenAsp <-> FGenAsp',      'psi12',       NA,
+     'ROccAsp <-> ROccAsp',      'theps1',      NA,
+     'REdAsp  <-> REdAsp',       'theps2',      NA,
+     'FOccAsp <-> FOccAsp',      'theps1',      NA,
+     'FEdAsp  <-> FEdAsp',       'theps2',      NA,
+     'FOccAsp <-> ROccAsp',      'theps24',     NA),
+     ncol=3, byrow=T)

> sem.dhp.2 <- sem(ram.dhp.2, R.dhp, N=329,
+     fixed.x=c('RParAsp', 'RIQ', 'RSES', 'FSES', 'FIQ', 'FParAsp'))

> summary(sem.dhp.2)

 Model Chisquare =  22.466   Df =  23 Pr(>Chisq) = 0.49228
 Goodness-of-fit index =  0.98643
 Adjusted goodness-of-fit index =  0.96755
 RMSEA index =  0   90% CI: (0, 0.044199)
 BIC =  -163.80
```

```
 Normalized Residuals
     Min.   1st Qu.    Median      Mean   3rd Qu.      Max.
-9.92e-01 -1.16e-01  6.29e-17 -2.46e-02  1.98e-01  6.77e-01

 Parameter Estimates
        Estimate Std Error  z value   Pr(>|z|)
gam1     0.160708  0.028662  5.60697 2.0590e-08 RGenAsp <--- RParAsp
gam2     0.307236  0.032455  9.46665 0.0000e+00     RGenAsp <--- RIQ
gam3     0.226074  0.032125  7.03732 1.9598e-12    RGenAsp <--- RSES
gam4     0.072527  0.037053  1.95738 5.0303e-02    RGenAsp <--- FSES
bet      0.204355  0.076737  2.66305 7.7435e-03 RGenAsp <--- FGenAsp
lamy     0.954089  0.051219 18.62761 0.0000e+00  REdAsp <--- RGenAsp
psi      0.278505  0.034585  8.05265 8.8818e-16 RGenAsp <--> RGenAsp
psi12    0.014493  0.054274  0.26703 7.8944e-01 FGenAsp <--> RGenAsp
theps1   0.337138  0.034399  9.80072 0.0000e+00 ROccAsp <--> ROccAsp
theps2   0.391574  0.034328 11.40669 0.0000e+00   REdAsp <--> REdAsp
theps24 -0.098785  0.031363 -3.14979 1.6339e-03 ROccAsp <--> FOccAsp

 Iterations =  26
```

References

Blau, P. M. & O. D. Duncan. 1967. *The American Occupational Structure*. New York: Wiley.

Bollen, K. A. 1989. *Structural Equations with Latent Variables*. New York: Wiley.

Bollen, K. A. & J. S. Long, eds. 1993. *Testing Structural Equation Models*. Newbury Park CA: Sage.

Duncan, O. D. 1975. *Introduction to Structural Equation Models*. New York: Academic Press.

Duncan, O. D., A. O. Haller & A. Portes. 1968. "Peer Influences on Aspirations: A Reinterpretation." *American Journal of Sociology* 74:119–137.

Fox, J. 1984. *Linear Statistical Models and Related Methods: With Applications to Social Research*. New York: Wiley.

Greene, W. H. 1993. *Econometric Analyis, Second Edition*. New York: Macmillan.

Jöreskog, K. G. 1973. A General Method for Estimating a Linear Structural Equation System. In *Structural Equation Models in the Social Sciences*, ed. A. S. Goldberger & O. D. Duncan. New York: Seminar Press pp. 85–112.

Judge, G. G., W. E. Griffiths, R. C. Hill, H. Lütkepohl & T.-C. Lee. 1985. *The Theory and Practice of Econometrics, Second Edition*. New York: Wiley.

Klein, L. 1950. *Economic Fluctuations in the United States 1921–1941*. New York: Wiley.

McArdle, J. J. 1980. "Causal Modeling Applied to Psychonomic Systems Simulation." *Behavior Research Methods and Instrumentation* 12:193–209.

McArdle, J. J. & R. P. McDonald. 1984. "Some Algebraic Properties of the Reticular Action Model." *British Journal of Mathematical and Statistical Psychology* 37:234–251.

McDonald, R. P. & W. Hartmann. 1992. "A Procedure for Obtaining Initial Values of Parameters in the RAM Model." *Multivariate Behavioral Research* 27:57–76.

Raftery, A. E. 1993. Bayesian Model Selection in Structural Equation Models. In *Testing Structural Equation Models*, ed. K. A. Bollen & J. S. Long. Newbury Park CA: Sage pp. 163–180.