January 3, 2000

Interpretation of R^2 Revisited

Karl G Jöreskog

In a previous note on this corner (*What is the Interpretation of R^2?*), I discussed the interpretation of R^2 for the equations of a *non-recursive system* and concluded that R^2 calculated from the structural equations does not have a clear interpretation and that, if one wants an interpretation similar to that of a regression equation, one must calculate R^2 from the reduced form equations rather than from the structural equations. The same argument applies to a *recursive system* but, under certain conditions, an R^2 calculated from a structural equation of a recursive system can be interpreted in a specific way. This note attempts to clarify this distinction.

Since the reduced form plays a fundamental role in the interpretation of R^2, we have added the reduced form in the standard output of LISREL. Thus, the standard output contains both the structural form and the reduced form and the R^2's calculated from each. This change has been made in December 1999 (Patch 7). Examples are given at the end of this note.

Previously, the reduced form could only be obtained in LISREL output format by putting `EF` (for indirect and total effects) on the `OU` line in a LISREL syntax file or on a `LISREL Output` or an `Options` line in a SIMPLIS syntax file. The reduced form was then given without any R^2's attached to the equations.

Consider a regression equation between a dependent variable y and a set of explanatory variables $\mathbf{x}' = (x_1, x_2, \ldots, x_q)$:

$$y = \alpha + \gamma_1 x_1 + \gamma_2 x_2 + \cdots + \gamma_q x_q + z \,, \tag{1}$$

or in matrix form

$$y = \alpha + \boldsymbol{\gamma}'\mathbf{x} + z \,, \tag{2}$$

where α is an intercept parameter, z is a random error term assumed to be uncorrelated with the explanatory variables, and $\boldsymbol{\gamma}' = (\gamma_1, \gamma_2, \ldots, \gamma_q)$ is a vector of coefficients to be estimated. As most textbooks on statistics or econometrics covering the topic of regression analysis will explain (see, *e.g.*, Goldberger, 1964), *the squared multiple correlation* also called *the coefficient of determination* is defined as

$$R^2 = 1 - \mathrm{Var}(z)/\mathrm{Var}(y) \,. \tag{3}$$

In practice, we may estimate R^2 by substituting the estimated variance of z for $\mathrm{Var}(z)$ and the estimated variance of y for $\mathrm{Var}(y)$ in (3). For the calculation of R^2 there are several equivalent formulas. It is common practice to provide an R^2 for every linear relationship estimated and LISREL has been doing so from LISREL 5.

The usual interpretation of R^2 is as the relative amount of variance of the dependent variable y explained or accounted for by the explanatory variables $x_1, x_2, \ldots, x_q$. For example, if $R^2 = 0.762$ we say that the explanatory variables "explains" 76.2% of the variance of y.

The main point here is that this interpretation of R^2 is not valid if we use definition (3) for relationships in a recursive or non-recursive system.

To explain this let $\mathbf{y} = (y_1, y_2, \ldots, y_p)$ be a set of jointly dependent (endogenous) variables and let $\mathbf{x} = (x_1, x_2, \ldots, x_q)$ be a set of independent (exogenous) variables. Consider a model of the form

$$\mathbf{y} = \boldsymbol{\alpha} + \mathbf{B}\mathbf{y} + \boldsymbol{\Gamma}\mathbf{x} + \mathbf{z} \,, \tag{4}$$

where $\boldsymbol{\alpha} = (\alpha_1, \alpha_2, \ldots, \alpha_p)$ is a vector of intercept terms, $\mathbf{B}$ and $\boldsymbol{\Gamma}$ are matrices of coefficients to be estimated, and $\mathbf{z} = (z_1, z_2, \ldots, z_p)$ is a vector of error terms assumed to be uncorrelated with $\mathbf{x}$. The x-variables are the only explanatory (exogenous) variables. The matrix $\mathbf{I} - \mathbf{B}$ is assumed to be non-singular. There are no latent variables in the model.

In scalar notation, equation (4) is

$$y_i = \alpha_i + \beta_{i1}y_1 + \beta_{i2}y_2 + \cdots + \beta_{ip}y_p + \gamma_{i1}x_1 + \gamma_{i2}x_2 + \cdots + \gamma_{iq}x_q + z_i, \quad i = 1, 2, \ldots, p \tag{5}$$

where some of the β's and γ's may be zero. If $\beta_{im} = 0$, y_i does not depend on y_m and if $\gamma_{in} = 0$, y_i does not depend on x_n. In general, (5) is not a regression equation because z_i may be correlated with the y-variables appearing on the right side of the equation.

For this equation to be identified, some of the β's and γ's must be zero. A simple neccessary but not sufficient condition for identification is the following. *For each y-variable included on the right side of (5) there must be at least one x-variable excluded from the same equation.* This is the so called order condition. There is also a rank condition which is both necessary and sufficient for identification (see, *e.g.*, Goldberger, 1964, p. 316), but this is difficult to apply in practice.

Consider the following simple non-recursive system with $p = 2$ and $q = 3$ (for simplicity, I assume that all but one of the coefficients are 1):

$$y_1 = y_2 + x_1 + z_1 \tag{6}$$

$$y_2 = 0.5y_1 + x_2 + x_3 + z_2 \tag{7}$$

It is obvious that the order condition is satisfied.

Based on the structural equations (6) and (7)

$$R_1^2 = 1 - \mathrm{Var}(z_1)/\mathrm{Var}(y_1) \tag{8}$$

for the first equation, and

$$R_2^2 = 1 - \mathrm{Var}(z_2)/\mathrm{Var}(y_2) \tag{9}$$

for the second equation.

The problem is that z_1 in (6) is not uncorrelated with y_2 appearing in that equation. So (6) is not a regression equation as in (1). To put it differently, the right side of (6) is not the conditional expectation of y_1 for given y_2 and x_1. Therefore, we cannot divide the variance of y_1 between z_1 and the other variables on the right side of (6). Also, this definition includes all of the variance of y_2 in the calculation of $\mathrm{Var}(y_1)$ although some of the variance of y_2 is due to error. The variance of y_1 depends on the variance of y_2 and vice versa. The interpretation of R_1^2 is therefore unclear. The same kind of argument applies to the second equation as well.

A better definition of R^2 can be obtained by using the *reduced form*, see Jöreskog & Sörbom (1996a, pp. 143–145). The reduced form is obtained by first noting that (4) can be written

$$(\mathbf{I} - \mathbf{B})\mathbf{y} = \boldsymbol{\alpha} + \boldsymbol{\Gamma}\mathbf{x} + \mathbf{z}\ , \tag{10}$$

and then premultiplying this by $(\mathbf{I} - \mathbf{B})^{-1}$. This gives the reduced form as

$$\mathbf{y} = (\mathbf{I} - \mathbf{B})^{-1}\boldsymbol{\alpha} + (\mathbf{I} - \mathbf{B})^{-1}\boldsymbol{\Gamma}\mathbf{x} + \mathbf{z}^*\ , \tag{11}$$

where $\mathbf{z}^* = (\mathbf{I} - \mathbf{B})^{-1}\mathbf{z}$. This equation is the multivariate regression (implied by the model) of $\mathbf{y}$ on $\mathbf{x}$. Since $\mathbf{z}^*$ is a linear combination of $\mathbf{z}$, $\mathbf{z}^*$ is uncorrelated with $\mathbf{x}$.

For the i-th equation in (11), R^2 can be defined as

$$R_i^{*2} = 1 - \mathrm{Var}(z_i^*)/\mathrm{Var}(y_i)\ . \tag{12}$$

This R_i^{*2} can be interpreted as the relative variance of y_i explained or accounted for by all explanatory variables jointly.

For the simple example, the reduced form is

$$y_1 = 2x_1 + 2x_2 + 2x_3 + z_1^* \tag{13}$$

$$y_2 = x_1 + 2x_2 + 2x_3 + z_2^* \tag{14}$$

where z_1^* and z_2^* are linear combinations of z_1 and z_2 and therefore uncorrelated with all the explanatory variables. Hence,

$$R_1^{*2} = 1 - \mathrm{Var}(z_1^*)/\mathrm{Var}(y_1) \tag{15}$$

$$R_2^{*2} = 1 - \mathrm{Var}(z_2^*)/\mathrm{Var}(y_2) \tag{16}$$

and each R^{*2} can be interpreted as the relative variance of the dependent variable explained or accounted for by all three x-variables jointly.

To simplify the calculations I assume that x_1, x_2, x_3, z_1, and z_2 are independent each with a variance of 1. From the reduced form it then follows that $R_1^{*2} = 0.60$ and $R_2^{*2} = 0.64$. If we calculate R^2 from the structural equations we obtain $R_1^2 = 0.95$ and $R_2^2 = 0.93$, but these R^2's have no clear interpretation.

To verify these results run the following SIMPLIS command file

```
Test of Small Non-Recursive SEM
Observed Variables: Y1 Y2 X1 X2 X3
Covariance Matrix
20 16 14 2 1 1 2 2 0 1 2 2 0 0 1
Sample Size: 101
Relationships
Y1 = Y2 X1
Y2 = Y1 X2 X3
End of Problem
```

This gives the following results:

```
  Structural Equations

Y1 = 1.00*Y2 + 1.00*X1, Errorvar.= 1.00 , R^2 = 0.95
    (0.032)   (0.11)              (0.16)
     31.14     9.39               6.36

Y2 = 0.50*Y1 + 1.00*X2 + 1.00*X3, Errorvar.= 1.00 , R^2 = 0.93
    (0.039)   (0.13)    (0.13)              (0.21)
     12.71     7.79      7.79               4.70

  Reduced Form Equations

Y1 = 2.00*X1 + 2.00*X2 + 2.00*X3, Errorvar.= 8.00, R^2 = 0.60
    (0.26)    (0.24)    (0.24)
     7.79      8.32      8.32

Y2 = 1.00*X1 + 2.00*X2 + 2.00*X3, Errorvar.= 5.00, R^2 = 0.64
    (0.19)    (0.21)    (0.21)
     5.34      9.39      9.39
```

Here, the last two R^2's are the only ones that have a clear interpretation, namely x_1, x_2, and x_3 explain 60% of the variance of y_1 and 64% of the variance of y_2. Note that the R^2's for the structural equations grossly overestimates these R^2's.

Next consider a simple recursive system with $p = 3$ and $q = 1$:

$$y_1 = x_1 + z_1 \,, \tag{17}$$

$$y_2 = y_1 + x_1 + z_2 \,, \tag{18}$$

$$y_3 = y_1 + y_2 + x_1 + z_3 \,, \tag{19}$$

where x_1, z_1, z_2, z_3, are mutually uncorrelated. From this assumption it follows that z_2 is uncorrelated with y_1 and that z_3 is uncorrelated with both y_1 and y_2, so that all three equations are regression equations. Therefore, R^2 for each equation can be interpreted as the proportion of variance of the left hand variable accounted for by the right hand variables. Assuming that x_1, z_1, z_2, z_3 all have variance one, the three R^2's calculated from these equations are 0.50, 0.83, 0.95. These R^2's can be interpreted as follows: x_1 explains 50% of the variance of y_1. x_1 and y_1 explain 83% of the variance of y_2. x_1, y_1, and y_2 explain 95% of the variance of y_3. The general conditions under which this kind of interpretation applies is that all elements in the diagonal and above the diagonal of $\mathbf{B}$ must be fixed zeroes and $\mathbf{\Psi}$, the covariance matrix of $\mathbf{z}$, must be diagonal.

The reduced form equations are

$$y_1 = x_1 + z_1^* \,, \tag{20}$$

$$y_2 = 2x_1 + z_2^* \,, \tag{21}$$

$$y_3 = 4x_1 + z_3^* \,, \tag{22}$$

where $z_1^* = z_1$, $z_2^* = z_1 + z_2$, $z_3^* = z_1 + z_2 + z_3$. R^2 calculated from these equations has the interpretation as the proportion of variance of the left hand variable accounted for by x_1. The three R^2's calculated from these equations are 0.50, 0.67, 0.73. Again, the R^2's calculated from the structural form overestimates the R^2's calculated from the reduced form.

To verify the results for the simple recursive model, run the following SIMPLIS command file:

```
Test of Small Recursive SEM
Observed Variables: Y1 Y2 Y3 X1
Covariance Matrix
2 3 6 6 11 22 1 2 4 1
Sample Size: 101
Relationships
Y1 = X1
Y2 = Y1 X1
Y3 = Y1 Y2 X1
End of Problem
```

This gives the following results

```
         Structural Equations

       Y1 = 1.00*X1, Errorvar.= 1.00 , R^2 = 0.50
           (0.10)              (0.14)
            9.95                7.04
```

```
Y2 = 1.00*Y1 + 1.00*X1, Errorvar.= 1.00 , R^2 = 0.83
    (0.10)    (0.14)              (0.14)
     9.95      7.04                7.04

Y3 = 1.00*Y1 + 1.00*Y2 + 1.00*X1, Errorvar.= 1.00 , R^2 = 0.95
    (0.14)    (0.10)    (0.17)              (0.14)
     7.04      9.95      5.74                7.04


  Reduced Form Equations

Y1 = 1.00*X1, Errorvar.= 1.00, R^2 = 0.50
    (0.10)
     9.95

Y2 = 2.00*X1, Errorvar.= 2.00, R^2 = 0.67
    (0.14)
     14.07

Y3 = 4.00*X1, Errorvar.= 6.00, R^2 = 0.73
    (0.25)
     16.25
```

All of the above applies to *latent* recursive and non-recursive models as well. Replacing $\mathbf{y}$ by $\boldsymbol{\eta}$, $\mathbf{x}$ by $\boldsymbol{\xi}$, and $\mathbf{z}$ by ζ, we get the structural equation model in LISREL:

$$\boldsymbol{\eta} = \boldsymbol{\alpha} + \mathbf{B}\boldsymbol{\eta} + \boldsymbol{\Gamma}\boldsymbol{\xi} + \zeta \,. \tag{23}$$

As a third example, consider the Hypothetical Model on pp. 133–135 in Jöreskog & Sörbom (1996b). For example, run the following SIMPLIS command file:

```
Hypothetical Model
Observed Variables: Y1-Y4 X1-X7
Correlation Matrix from File EX17.COV
Sample Size: 100
Latent Variables: Eta1 Eta2 Ksi1-Ksi3
Relationships
    Eta1 = Eta2 Ksi1 Ksi2
    Eta2 = Eta1 Ksi1 Ksi3
Let the Errors of Eta1 and Eta2 Correlate

    Y1 = 1*Eta1
    Y2 = Eta1
    Y3 = 1*Eta2
    Y4 = Eta2

    X1 = 1*Ksi1
    X2 X3 = Ksi1
    X4 = 1*Ksi2
    X3 X5 = Ksi2
```

```
    X6 = 1*Ksi3
    X7 = Ksi3
End of Problem
```

This gives the following results:

```
         Structural Equations

     Eta1 = 0.54*Eta2 + 0.21*Ksi1 + 0.50*Ksi2, Errorvar.= 0.49 , R^2 = 0.84
           (0.056)     (0.15)      (0.15)                (0.13)
            9.53        1.39        3.35                  3.83

     Eta2 = 0.94*Eta1 - 1.22*Ksi1 + 1.00*Ksi3, Errorvar.= 0.13  , R^2 = 0.97
           (0.18)      (0.12)      (0.15)                (0.078)
            5.25       -10.05       6.57                  1.70

         Reduced Form Equations

     Eta1 =  - 0.90*Ksi1 + 1.00*Ksi2 + 1.08*Ksi3, Errorvar.= 1.83, R^2 = 0.38
              (0.43)      (0.34)      (0.22)
              -2.09        2.92        4.81

     Eta2 =  - 2.06*Ksi1 + 0.94*Ksi2 + 2.01*Ksi3, Errorvar.= 1.75, R^2 = 0.63
              (0.52)      (0.40)      (0.27)
              -3.99        2.32        7.49
```

Here the two R^2's of 0.84 and 0.97 based on the structural equations have no clear interpretation. The other two R^2's of 0.38 and 0.63 based on the reduced form mean that `Ksi1`, `Ksi2`, and `Ksi3` explain 38% of the variance of `Eta1` and 63% of the variance of `Eta2`.

References

Goldberger, A.S. (1964) *Econometric theory.* New York: Wiley.

Jöreskog & Sörbom (1996a) *LISREL 8: User's Reference Guide.* Chicago: Scientific Software International.

Jöreskog & Sörbom (1996b) *LISREL 8: Structural Equation Modeling with the SIMPLIS Command Language.* Chicago: Scientific Software International.