What is the interpretation of R^2?

Karl G. Jöreskog

October 2, 1999

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

$$y = \alpha + \gamma_1 x_1 + \gamma_2 x_2 + ... + \gamma_q x_q + z, \quad (1)$$

or in matrix form

$$y = \alpha + \gamma' \mathbf{x} + z, \quad (2)$$

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

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

In practice, we may estimate R^2 by substituting the estimated variance of z for Var(z) and the estimated variance of y for 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 version 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 $x_1, x_2,..., x_q$. 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 non-recursive system. For this reason, the definition of R^2 has been changed in LISREL 8.30, with the release of the released August 1999 (Patch 3) version.

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

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

where $\boldsymbol{\alpha} =(\alpha_1, \alpha_2, ..., \alpha_p)$ is a vector of intercept terms, $\mathbf{B}$ and Γ are matrices of coefficients to be estimated, and $\mathbf{z} =(z_1, z_2, ..., z_p)$ is a vector of error terms assumed to be uncorrelated with $\mathbf{x}$. The matrix $\mathbf{I} - \mathbf{B}$ is assumed to be non-singular. There are no latent variables in the model. Suppose the system is non-recursive so that the equations cannot be ordered in such a way that $\mathbf{B}$ is sub-diagonal (see Jöreskog & Sörbom, 1996a, pp 143-145).

In scalar notation, equation (4) is

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

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. 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 for example, Goldberger, 1964, p.316), but this is difficult to apply in practice.

Consider the following simple example with p = 2 and q = 3:

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

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

It is obvious that the order condition is satisfied.

The previous versions of LISREL (prior to August 1999) used the definition

$$\mathrm{R}_1^2 = 1 - \mathrm{Var}(z_1)/\mathrm{Var}(y_1) \tag{8}$$

for the first equation, and

$$\mathrm{R}_2^2 = 1 - \mathrm{Var}(z_2)/\mathrm{Var}(y_2) \tag{9}$$

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 Var(y_1) although some of the variance of y_2 is due to error. The variance of y_1 depends on the variance 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 for non-recursive systems 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 as

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

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

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

where $\mathbf{z}^* = (\mathbf{I} - \mathbf{B})^{-1}\, \mathbf{z}$. This equation is the multivariate regression (as 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 $\mathbf{x}$.

We can now define the new R_i^{*2} for the *i*-th equation in (11) as

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

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

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

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

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 follows that $R_1^{*2} = 0.60$ and $R_2^{*2}=0.64$. With previous definitions we obtain $R_1^2 = 0.95$ and $R_2^2 = 0.93$. We should therefore expect large differences in R^2 between the previous and the current version of LISREL.

To verify these results run the following SIMPLIS command file :

```
Test of Small 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:

```
Y1 = 1.00*Y2 + 1.00*X1, Errorvar.= 1.00 , R2 = 0.60
    (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 , R2 = 0.64
    (0.039)   (0.13)    (0.13)             (0.21)
     12.71     7.79      7.79               4.70
```

The previous version (prior to August 1999) of LISREL gave the following results:

```
Y1 = 1.00*Y2 + 1.00*X1, Errorvar.= 1.00 , R2 = 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 , R2 = 0.93
    (0.039)   (0.13)    (0.13)             (0.21)
     12.71     7.79      7.79               4.70
```

Note that parameter estimates, standard errors, and t-values are all the same. Only R^2 is different. The previous version overestimates the strength of the relationships.

All of the above applies to *latent* non-recursive models as well. Replacing **y** by η, **x** by ξ, and **z** by ζ, we get the structural equation model in LISREL:

$$\eta = \alpha + \mathbf{B}\,\eta + \Gamma\,\xi + \zeta. \tag{17}$$

The R^2s for these structural equations will also be different if **B** is not subdiagonal.

As a second example, consider the Hypothetical Model on pp. 133-135 in Jöreskog & Sörbom (1996b). For example, run the following SIMPLIS command file (adapted from the file EX17A.SPL in the SPLEX subdirectory):

```
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

!LISREL Output: RS MI SC EF WP
End of Problem
```

This gives the following results:

```
Eta1 = 0.54*Eta2 + 0.21*Ksi1 + 0.50*Ksi2, Errorvar.= 0.49 , R2 = 0.38
      (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 , R2 = 0.63
      (0.18)      (0.12)      (0.15)                (0.078)
       5.25       -10.05       6.57                  1.70
```

The previous version (prior to August 1999) of LISREL gave the following results:

```
Eta1 = 0.66*Eta2 + 0.14*Ksi1 + 0.32*Ksi2, Errorvar.= 0.15 , R2 = 0.84
      (0.069)     (0.10)      (0.097)               (0.040)
       9.53        1.39        3.35                  3.83

Eta2 = 0.76*Eta1 - 0.65*Ksi1 + 0.54*Ksi3, Errorvar.= 0.027 , R2 = 0.97
      (0.14)      (0.064)     (0.082)               (0.016)
       5.25       -10.05       6.57                  1.70
```

Again note that parameter estimates, standard errors, and t-values are all the same. Only R^2 is different. The previous version overestimates the strength of the relationships.

References

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

Jöreskog, K.G. & Sörbom, D. (1996a) LISREL8: *User's Reference Guide*. Chicago: Scientific Software International.

Jöreskog, K.G. & Sörbom, D. (1996b) LISREL8: *Structural Equation Modeling with the SIMPLIS Command Language*. Chicago: Scientific Software International.