Automated Path Tracing for General Linear Models

William Dardick **Jeffrey R. Harring**
University of Maryland

Monte Carlo simulations in statistics are computer experiments involving random sampling from known probability distributions to study properties of statistical methods. As the relations among variables become increasingly complex, the method of generating data under imposed model and distributional conditions becomes progressively more complicated. In this article, we algebraically derive a kernel for calculating direct causal effects in a path model for the univariate general linear model (GLM), specifically discussing regression models and extensions such as analysis of variance. A rationale is provided for decisions made for dealing with multiple unknown parameters and predictor correlation, where all unknown parameters are assumed to be equal. Separate from the methodological value is the real world application involving solving for a full correlation matrix, sample size, and power analysis in a GLM framework.

Monte Carlo simulation is an invaluable and versatile tool that can be used in situations where the researcher wishes to examine particular effects under strictly-controlled distributional or parameter-constrained conditions. For example, in regression with correlated normally distributed errors, the properties of the generalized least squares (GLS) estimator can be derived in a straightforward manner when the covariance matrix, Var(**y**), is known. However, when Var(**y**) is unknown, the impact of substituting an estimate of Var(**y**) into the GLS equations is unclear. Large sample properties can be ascertained for the case when the sample size tends to infinity, yet the only way to get information about the estimator's small sample behavior is to conduct a statistical simulation.

Monte Carlo simulation studies involve generating data under a variety of model-specific and distributional misspecification with the explicit purpose of investigating such phenomena as (*i*) the behavior of statistical estimators under violations of the assumptions underlying their use (Hutchinson & Bandelos, 1997); (*ii*) the comparisons of the accuracy and efficiency of different methods and/or computational algorithms that have been designed to do the same thing; and (*iii*) evaluation of new statistical estimators (Harwell, Stone, Hsu, & Kirisci,1996). Typical outcomes (dependent variables) commonly studied in Monte Carlo simulations include Type I and II error rates (probability coverage for confidence intervals), bias and variance of estimators, and power functions of hypothesis tests. Although studies of error rates and parameter estimate bias seem to constitute the majority of simulation research (Hutchinson & Bandelos), there are numerous other potentially beneficial outcomes that could be explored, yet are tailored to reflect nuances of a particular methodology (i.e., structural equation modeling, item response theory). Finally, the literature seems replete with studies that have primarily focused on the impact and reporting of independent (manipulated) variables have on the dependent variable(s); yet, less attention has been devoted to the process of data generation for a particular model.
Here, the model reflects a conceptual or theoretical representation of the phenomenon of interest. The task for simulation researchers is to transform the conceptual model into a mathematical model, which serves as the basis for data generation. For example, in an investigation to determine the least biased estimator among four competitors of the population multiple coefficient of determination, $\hat{R}^2$, a multiple regression model would need to be constructed with known population value P^2. Because the formulae to compute the different R^2 values are dependent upon sample size and number of predictor variables, at a minimum, these two quantities should be included as independent variables in the simulation design. However, the overall R^2 value is determined by both the correlation between each individual predictor and the outcome as well as the intercorrelation among the predictors. It is not clear how these correlations should be specified to obtain a particular known overall R^2. Additionally, and perhaps more importantly for multiple regression simulations, it is not clear how the regression coefficients should be specified given particular multicollinearity and overall R^2.

It is within the context of multiple regression, and more broadly the general linear model (GLM), that we propose an algorithmic path tracing scheme to automatize the data generation process. The primary objectives and subsequent chronology of this article are as follows: (*i*) to algebraically derive the key relationships between R^2, the matrix of predictor correlations, and the number of predictors; (*ii*) to use the

algebraic formulas to solve for unknown coefficients, where all unknown coefficients are assumed to be equal; (*iii*) to provide a general model in matrix formulas to solve for GLM path models given any combination of known values; and (*iv*) to illustrate the utility of the method through several worked examples using a self-generated computational tool.

Path Tracing for Regression Models

Multiple regression requires a basic understanding of sample statistics (e.g., *n*, mean, and variance), standardized variables, correlation (Pedhazur & Pedhazur-Schmelkin, 1991), and partial correlation (Cohen, Cohen, West, & Aiken, 2003). In standard-score form, the multiple regression equation is defined as:

$$\hat{z}_{y_i} = \beta_1 z_{x_{1i}} + \beta_2 z_{x_{2i}} + \cdots + \beta_p z_{x_{pi}}$$
$$\hat{z}_{y_i} = \sum_{k=1}^{p} \beta_k z_{x_k} \quad (1)$$

where β_k are the standardized regression coefficients. In the language of path analysis, standardized partial regression coefficients, or beta weights, in multiple regression problems are called path coefficients for short (Loehlin, 1998). When all variables are measured, one can solve for these paths as you would for beta coefficients.

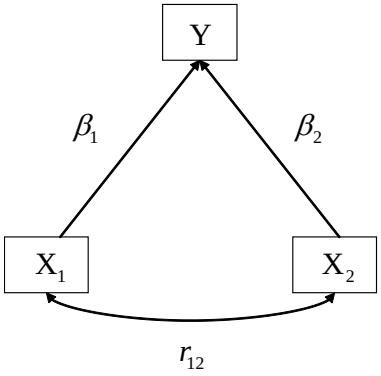


Figure 1. Path diagram of a regression model.

Figure 1 is a simple path diagram of a regression model with 2 independent variables (IV), or *X*, and one dependent variable, (DV), or *Y*. Using Wright's rules (Wright, 1960), or through manipulation of beta equations (Loehlin, 1998), the coefficient of determination, R^2, implicit in the path diagram can be solved for as follows: $R^2 = \beta_1^2 + \beta_2^2 + 2\beta_1\beta_2 r_{12}$ where, r_{12} represents the correlation between X_1 and X_2, and β_1 and β_2 are the standardized partial regression coefficients. Extending this same model to 3 IVs results in:

$$R^2 = \beta_1^2 + \beta_2^2 + \beta_3^2 + 2\beta_1\beta_2 r_{12} + 2\beta_1\beta_3 r_{13} + 2\beta_2\beta_3 r_{23} \quad (2)$$

where β_1, β_2, and β_3 are the path coefficients and r_{12}, r_{13}, and r_{23} are the correlations between the IVs. While the expression for R^2 becomes progressively more convoluted as the number of IVs increases, a relatively straightforward pattern begins to emerge. A generalized pattern for an expression of R^2 with *p* IVs is:

$$R^2 = \sum_{i=1}^{p} \beta_i^2 + 2\sum_{i=1}^{p-1} \sum_{j=i+1}^{p} \beta_i \beta_j r_{ij} \quad (3)$$

Two kernels of patterns are present in the generalized expression for R^2 in Equation 3. The squared path coefficients, $R_v^2 = \beta_1^2 + \beta_2^2 + \ldots + \beta_p^2 = \sum_{i=1}^{p} \beta_i^2$, can be interpreted as the variance component of the model where R_v^2 is the portion of the coefficient of determination derived from the variance components of the model and where $i = 1, \ldots, p$ indexes the coefficient corresponding to the i^{th} IV.

The portion of the coefficient of determination attributable to covariance defines the second kernel: $R_c^2 = 2\beta_1\beta_2 r_{12} + 2\beta_1\beta_3 r_{13} + \ldots + 2\beta_i\beta_j r_{ij} = 2\sum_{i=1}^{p-1} \sum_{j=i+1}^{p} \beta_i \beta_j r_{ij}$.

Traditionally, R^2 is written as:

$$R^2 = \sum_{i=1}^{p} r_{yx_i} \beta_i \quad (4)$$

where r_{yx_i} represents the correlation between the DV and i^{th} IV and β_i is the standardized beta coefficient. However, one could rewrite Equation 4 without the DV and IV correlations and; instead, use only the correlations amongst the IVs:

$$R^2 = \sum_{i=1}^{p} \sum_{j=1}^{p} \beta_i r_{ij} \beta_j \quad (5)$$

where β_i and β_j are the standardized beta coefficient and r_{ij} is the correlation between two IV variables x_i and x_j over $i = 1, \ldots, p$ and $j = 1, \ldots, p$ for all i and j.

Solving for Beta

In order to solve an arbitrary regression equation with multiple unknown beta coefficients, a theoretical decision needs to be made related to how to deal with unknown path information. Setting all unknown direct paths equal to one another cleans up the model and simplifies the result. In doing so, the issue of not having measurements for all variables can be avoided. When all of the standardized regression coefficients, β's, are equal, Equations 3 and 5 simplify. As an example, imagine the case where $p = 4$ and with all of the variances set equal to one another. The expression for R^2 is written as:

$$R^2 = \beta_1^2 + \beta_2^2 + \beta_3^2 + \beta_4^2 + 2\beta_1\beta_2 r_{12} + 2\beta_1\beta_3 r_{13} + 2\beta_1\beta_4 r_{14} + 2\beta_2\beta_3 r_{23} + 2\beta_2\beta_4 r_{24} + 2\beta_3\beta_4 r_{34}.$$

The variance portion of the model simplifies to $4\beta^2$ because all β values are equal. The equation can now be expressed as:

$$R^2 = 4\beta^2 + 2\beta_1\beta_2 r_{12} + 2\beta_1\beta_3 r_{13} + 2\beta_1\beta_4 r_{14} + 2\beta_2\beta_3 r_{23} + 2\beta_2\beta_4 r_{24} + 2\beta_3\beta_4 r_{34}.$$

Further assuming all correlations amongst the IVs are the same, yields an even more simplified version:

$$R^2 = 4\beta^2 + 2r(\beta_1\beta_2 + \beta_1\beta_3 + \beta_1\beta_4 + \beta_2\beta_3 + \beta_2\beta_4 + \beta_3\beta_4).$$

Because all path coefficients β are equal, the expression for R^2 simplifies to:

$$\begin{aligned} R^2 &= 4\beta^2 + 2r(6(\beta^2)) \\ &= 4\beta^2 + 12r\beta^2 \\ &= \beta^2(4 + 12r) \end{aligned}$$

For a specific value or R^2, the path coefficient is the positive root of the quadratic:

$$\beta^2 = \frac{R^2}{(4+12r)}, \text{ namely } \beta = \sqrt{\frac{R^2}{(4+12r)}}$$

While the result above is specific for a regression model with 4 predictors, solving for β can be easily generalized for any number of predictors given the restrictions that the coefficients are identical and the correlations between the IVs are also the same:

$$R^2 = \beta^2 p + \beta^2 r[p(p-1)] \tag{6}$$

Solving the expression for R^2 in Equation 6 for β, results in:

$$\beta = \sqrt{\frac{R^2}{p + r[p(p-1)]}} \tag{7}$$

where β is the path coefficient of interest, and where p denotes the number of independent variables in the model. The IV correlations dictate the number of terms in the model. There are $p(p-1)/2$ correlations. Because r is used for each possible direction, the n for the covariance portion is simply $p(p-1)$.

Solving for the Full Correlation Matrix

An unexpected benefit of deriving the beta coefficients in the regression model is that by doing so, it provides insight into specifying the full correlation matrix of the variables; including a reasonable estimate of what the correlations between IVs and DV might be. The expression for β in Equation 7 can be used to build the correlation from the full set of coefficients. This can be accomplished in the following manner:

$$r_{yj} = \beta + \beta \cdot (p-1) \cdot r_{jk} \tag{8}$$

where r_{yj} is the correlation between the DV and the j^{th} IV. Each bivariate correlation r_{jk} is set to be the same overall correlation r and β is the path coefficient solved for using Equation 7. All correlations will be the same under the null assumption of no difference. To see how these various pieces fit into a cohesive framework, consider the following example.

Example 1

Consider the following 2 IV regression problem in Figure 2 where the coefficient of determination is set at 0.50 and the correlation between the two predictors are specified as 0.30. Using Equation 7 we can derive the standardized partial regression coefficients:

$$\beta = \sqrt{\frac{R^2}{p + r[p(p-1)]}} = \frac{.5}{2 + .3\left[2(2-1)\right]} = 0.438.$$

The value of the path coefficient, 0.438, can be used to verify the R^2 value with Equation 1: $R^2 = \beta_1^2 + \beta_2^2 + 2\beta_1 \beta_2 r_{12}$

$= 0.438^2 + 0.438^2 + (0.438 * 0.438 * 0.30)$

$= 0.50$

Equation 8 can be used to compute the correlations between the two IVs and the DV as follows:

$$r_{yj} = \beta + \beta \cdot (p-1) \cdot r_{jk} = 0.438 + 0.438 \cdot (2-1) \cdot 0.30 = 0.57.$$

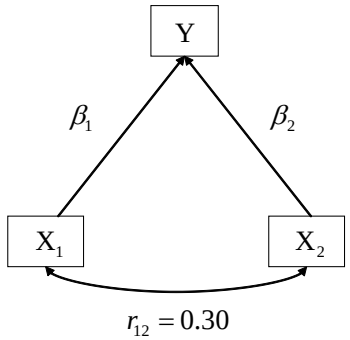


Figure 2. Two IV regression model

Example 2

As a second illustration, Fan, Felsővályi, Sivo, and Keenan (2001) demonstrated how one might compare different R^2 shrinkage formulae in regression analysis through simulation. One of the manipulated conditions that Fan et al. varied was the number of independent variables. For eight IVs with a constant inter-correlation of 0.30 and a population,R^2 = 0.50, the standardized partial regression weights and the complete data correlation matrix can be computed using Equations 1, 7, and 8 as follows:

$$\beta = \sqrt{\frac{R^2}{p + r[p(p-1)]}} = \frac{0.50}{8 + 0.30[8(8-1)]} = 0.142$$

Again, this can be checked with Equation 1:

$$R^2 = \beta^2 p + \beta^2 r[p(p-1)] = 0.142^2 * 8 + 0.142^2 * 0.30[8(8\text{-}1)] = 0.50$$

Correlations between any IV and the DV can be computed as follows:

$$r_{yj} = \beta + \beta \cdot (p-1) \cdot r_{jk} = 0.142 + 0.142 \cdot (8-1) \cdot 0.30 = 0.4402.$$

These results coincide with the results reported by Fan et al.; however, in their example, no explanation was provided as to how the data correlation matrix (which must necessarily be computed to generate the raw data for the simulation) was created. In contrast, after making a few initial assumptions, we demonstrate how this can be accomplished in a straightforward manner.

These two examples have application in Monte Carlo simulations where a correlation matrix holding to necessary parameters needs to be imputed to evaluate results. Instead of having to path trace the model for each instance, a simple calculation can occur to iterate over a prescribed number of factors in a simulation model. Factors such as number of variables, overall R^2 values, and predictor intercorrelations can be easily imputed as manipulated factors into simulation designs.

Generalizing the Equation

In many cases, a researcher may have *a priori* knowledge of some parameter values in the model (i.e., standardized partial regression coefficients or predictor correlations). Of practical importance, it would be most valuable to allow any path in the model to be set by the researcher. That is, instead of assigning direct paths to be equal, allowing for separate specification of model parameters would increase the flexibility and utility of the overall approach. Whether theoretically-based or driven by some other compelling rationale, any number of parameters can be set to desired values. Implications for this added flexibility will be explicated in the following example.

Example 3

Again, consider the expression for R^2 from a regression model with three IVs:

$$R^2 = \beta_1^2 + \beta_2^2 + \beta_3^2 + 2\beta_1\beta_2 r_{12} + 2\beta_1\beta_3 r_{13} + 2\beta_2\beta_3 r_{23}.$$

The equation can be formulated in a similar manner as before, but now allowing for parameter difference to exist. All unknown values are again set to be equal. Using the subscript notation "*k*" and "*u*" for known and unknown, respectively, an equation with one known IV and two unknown IVs (three total IVs) could be written as:

$$R^2 = \beta_k^2 + \beta_u^2 + \beta_u^2 + 2\beta_k\beta_u r_{ku} + 2\beta_k\beta_u r_{ku} + 2\beta_u\beta_u r_{uu},$$

where β_k is the known or hypothesized standardized coefficient while β_u denotes the unknown coefficients. All unknown parameters β_u are assumed to be equal. In this arrangement, r, β_k, and R^2 are specified by the researcher and can be considered constants in a newly formulated quadratic equation as a function of β_u, which can subsequently be set equal to 0 and solved for β_k.

$$0 = (\beta_k^2 - R^2) + 2(2\beta_k\beta_u r_{ku}) + (2\beta_u^2 + 2\beta_u^2 r_{uu}) \quad (9)$$

Although factoring the quadratic could be used as a method to finding roots of the function, we advocate using the quadratic formula in order to automate the process:

$$\beta_k = \frac{-b \pm \sqrt{b^2 - 4ac}}{2a},$$

where *a*, *b*, and *c* are the coefficients of the quadratic in Equation 9:

$$a = (2b_u^2 + 2b_u^2 r_{uu}) = (2(1_u^2) + 2(1_u^2) r_{uu}) = 2 + 2 * r$$
$$b = 2(2b_k b_u r_{ku}) = 4 * b_k * 1 * r_{ku}$$
$$c = (b_k^2 - R^2)$$

This approach can be easily generalized to any number of known and unknown parameters.

Using the formulation in Equation 3 to have both known and unknown parameters, leads to the modification:

$$R^2 = \sum_{k=1}^{K} \beta_k{}^2 + \sum_{u=1}^{U} \beta_u^2 + 2r\sum_{k=1}^{K}\sum_{u=1}^{U}(\beta_k\beta_u) + 2r\sum_{k=1}^{K-1}\sum_{l=k+1}^{K}(\beta_k\beta_l) + 2r\sum_{u=1}^{U-1}\sum_{v=u+1}^{U}(\beta_u\beta_v) \quad (10)$$

Again, setting the Equation 10 equal to zero and solving yields estimated quadratic coefficients whose formulation can be expressed in terms of known and unknown information. Thus, the coefficient *a* is a function of only unknown entities:

$$a = \sum_{u=1}^{U} \beta_u^2 + 2r\sum_{u=1}^{U-1}\sum_{v=u+1}^{U}(\beta_u\beta_v) \quad (11)$$

while the coefficient of the linear term, *b*, is a function of unknown and known quantities:

$$b = 2r\sum_{k=1}^{K}\sum_{u=1}^{U}(\beta_k\beta_u) \quad (12)$$

Lastly, the constant coefficient, *c*, is a function of only known quantities:

$$c = \sum_{k=1}^{K} \beta_k{}^2 + 2r\sum_{k=1}^{K-1}\sum_{l=k+1}^{K}(\beta_k\beta_l) - R^2 \quad (13)$$

These general algebraic solutions solicit a matrix resolution to readily solve for robust extensions of values of predictor intercorrelations.

Matrix Representation

The linear equations above have been worked out as matrix algebra expressions. The matrix expressions solve for paths more robustly than expressed in the linear equations. The matrix equations easily permit the correlation between pairs of predictors to be different across pairs. This enhancement allows practitioners the additional flexibility to completely specify the correlation matrix corresponding to real data analytic conditions thought to prevail in substantive research. Furthermore, multiple standardized partial regression coefficients can be easily set.

In matrix form, the squared multiple correlation, R^2, can be written as:

$$R^2 = \mathbf{r}'_{yx}\boldsymbol{\beta}_p \quad (14)$$

where $\mathbf{r}'_{yx}$ is a 1 x p vector of correlations between the DV and IVs and $\boldsymbol{\beta}_p$ is a p x 1 vector of standardized regression coefficients. In Equation 14, $\mathbf{r}_{yx}$ can be re-expressed in terms of only the intercorrelations of the IVs:

$$\mathbf{r}'_{yx} = \boldsymbol{\beta}'_p \mathbf{R}_{xx} \tag{15}$$

where $\mathbf{R}_{xx}$ is a p x p matrix of IV correlations. Through substitution, R^2 can now be written as:

$$R^2 = \boldsymbol{\beta}'_p \mathbf{R}_{xx} \boldsymbol{\beta}_p \tag{16}$$

Pre-multiplying by the inverse of $\mathbf{R}_{xx}$ and taking the square root of the resultant expression leads to:

$$\sqrt{\mathbf{R}_{xx}^{-1} R^2} = \sqrt{\boldsymbol{\beta}'_p \boldsymbol{\beta}_p} \tag{17}$$

Setting all standardized coefficients to be equal effectively permits their representation as a scaled scalar:

$$\beta = \sqrt{R^2 \cdot \mathbf{R}_{xx}^{-1}} \tag{18}$$

The equation transforms to simple linear algebra as the summation of values in the correlation matrix of IVs when all correlations are equal (i.e., $r_{jk} = r$). Note this is identical to Equation 7.

The equations to solve for R^2 assume that one knows all parameters including standardized coefficients or appropriate correlations. However, if R^2 is known, which may be the case when used as an overall effect size measure, the matrix structure can be rearranged into known, mixed and unknown parameters just like the previous development.

When all the standardized partial regression coefficients are unknown, this model simplifies greatly. Prior information permits the use of known values (or good estimates of values). The model becomes more complex, but can be easily reorganized into the structures of what is known and unknown as well as the mixture of parameter types. Here, we start with Equation 16, set it equal to zero, and solve for the unknown coefficients:

$$\begin{aligned} R^2 &= \boldsymbol{\beta}'_p \mathbf{R}_{xx} \boldsymbol{\beta}_p \\ 0 &= \boldsymbol{\beta}'_p \mathbf{R}_{xx} \boldsymbol{\beta}_p - R^2. \end{aligned} \tag{19}$$

Of the p coefficients in $\boldsymbol{\beta}_p$, k of them are unknown and assumed to be equal. Solving for these coefficients is facilitated by setting the coefficients in Equation 16 to one. To get a better sense of the intricacies of specifying different parts of the model and their effects on finding the roots of the resulting quadratic, we present the matrix information in a transparent manner.

For illustrative purposes, assume a regression model with 4 IVs, 2 of which are known and 2 unknown. Figure 3 shows that the matrix multiplication in Equation 19, $\boldsymbol{\beta}'_p \mathbf{R}_{xx} \boldsymbol{\beta}_p$, is:

β	β	x	x	⨿	1	r	r	r	⨿	β
					r	1	r	r		β
					r	r	1	r		x
					r	r	r	1		x

Figure 3. Matrix multiplication for Equation 19.

In order to proceed, Figure 4 shows that the IV correlation matrix, $\mathbf{R}_{xx}$, will be partitioned into quadrants corresponding to the parameters that are known and those that are unknown. When all standardized coefficients are unknown, then the matrix equivalent to Equation 11 is: $a = \boldsymbol{\beta}'_u \mathbf{R}_{Q_4} \boldsymbol{\beta}_u$. When there is a mixture of known and unknown coefficients, then Equation 12 becomes: $b = \boldsymbol{\beta}'_p \mathbf{R}_{Q_2 Q_3} \boldsymbol{\beta}_p$.Lastly, when each standardized coefficient is known, Equation 13 generalizes to: $c = \boldsymbol{\beta}'_k \mathbf{R}_{Q_1} \boldsymbol{\beta}_k - R^2$. The matrix equivalents to Equations 11, 12, and 13 are the major developments. Illustrations for these matrix expressions can be found in Appendix[1].

1	r	r	r	Q1	Q2
r	1	r	r		
r	r	1	r	Q3	Q4
r	r	r	1		

Figure 4. Quadrants with parameters known and unknown.

To solve the above formulae for the standardized regression coefficients, insert 1's in for all unknown parameters, x's, known β_k values are set to a desired value, and r is set to the universal specified value (the correlation values can be adjusted as well). We solve for the two null solutions when the correlations between DV and IVs are positive and negative. It is assumed for the following example that only the positive root is essential.

Three examples, along with information on a companion SAS macro for this article are provided that demonstrate the utility of the automated path tracing capabilities from the methods previously outlined. The first example represents a situation in which all parameters are unknown. In the second example, holding all else constant, we add a mixture of known and unknown standardized coefficients. The final example builds off of the second example and adds fully pre-specified values for the intercorrelations of the IVs.

When all parameter values are unknown, we define the coefficient paths we wish to solve for to 1. Moreover, the IV correlation matrix is set to zero for all areas without known information (for the first example this corresponds to setting the elements in quadrants 1, 2, and 3 all to zero).

These examples are contrived to mimic situations that researchers might find in practice. There are several ways of thinking about the outcomes from the SAS macro used to automate this process. Two of the direct outcomes of the macro are a correlation matrix including the *a priori* estimation of correlations between IVs and DVs and sample size determination based on the beta coefficient or direct pathways in the model. There will be times that a researcher might want to use the direct outcome of this process to determine a reasonable upper bound to sample size. By determining the path in the null condition based on information provided by a researcher, it is expected that, all else being constant, some of those paths must be statistically significant. Alternatively, a researcher or methodologist might be more interested in the automated creation of a correlation matrix with specific factors to manipulate. For example this macro could be used as a sub-routine in a simulation similar to the one proposed by Fan et al. (2001), where the researcher wants to vary the R^2, the number of IVs, and the intercorrelations of IVs. This can also be used for estimates of paths as an alternative to maximum likelihood without the need to supply a sample size *a priori*.

The examples use a macro developed in SAS specifically designed for support of the ideas and mathematics developed in the current article[2]. The macro also provides the ability to manipulate known or unknown parameters (R^2 = R2 and β = b in the macro) while the IV intercorrelations (MC in the macro) can be changed as desired[3].

Starting with the first example, the coefficient of determination is set $R^2 = 0.50$ and initially $r = 0.30$ while all the paths are set to be unknown. In the second example, two of the paths are set to be known and all else remains the same. Finally, the intercorrelations of IV are estimated by the researcher beforehand from a known set of data while the remaining information from example two remains the same.

The macro called DMATRIX has been simplified for use in this article. Some important sections of the macro are discussed here prior to providing the example so that an interested researcher can follow along and try some examples of their own.

The researcher has three options for creating the IV corrections. One can simply supply the main SAS macro with an omnibus multicollinearity for all IVs. The researcher can also input data into the preparation macro called PREPDMATRIX. PREPDMATRIX is an optional MACRO which prepares SAS datasets and SAS correlations datasets for use with the main DMATRIX MACRO when the IV relationships are known by the researcher or approximated individually in a SAS correlation format. The first option using this macro takes raw data and transforms it into a correlation matrix for IVs. The second option for this macros uses an existing correlation matrix in SAS format[4]. The macro statements and structure to invoke this code, once the macro is run locally, is as follows::

%PREPDMATRIX (IV,GENMCL,FULLMCL);

In the above macro, IV is the number of independent variables in the dataset. The dataset used needs to be an existing dataset of IVs the researcher intends to use or the correlations of those IVs. If using a SAS dataset, match the name of the dataset to the GENMCL parameter. For example, if your raw SAS dataset has 4 variables and is called MYDATA, your code would look like:

%PREPDMATRIX (4, MYDATA, FULLMCL);

If the dataset is a correlation dataset in SAS, in the format SAS produces using the OUTP option in a Proc Corr statement, match the name of that matrix to the FULLMCL parameter. For example, if your correlation matrix in SAS has 4 variables and is called MYCORRMATRIX, your code would look like:

%PREPDMATRIX (4, GENML, MYCORRMATRIX);

The variable names need to be changed to read IV1 through IV#, where # is the last IV. The end result of this optional macro will be a correlation dataset called FULLMCL and used in the main macro

DMATRIX containing specified values for correlations amongst IVs. If you do not have your own data, you will not need to change the names in this macro and you can skip it entirely. It also should be noted that once the outcome matrix FULLMCL has been created, it can be manipulated to place any correlation values a researcher would choose if the researcher is familiar enough to copy over values in a SAS correlation dataset.

The main macro, DMATRIX, is a non-iterative method to calculate path values as described in the above equations. The macro statements and structure to invoke the DMATRIX code, once the macro is run locally, is as follows:

%DMATRIX (VN, R2, MCL, FULLMCL, NB, SB);

Two of the above parameters are the same as in the PREPDMATRIX code. IV is still the total number of IVs. If importing an existing dataset with known values, the parameter FULLMCL can be used from the PREPDMATRIX macro. R2 is the overall R^2 value set by the researcher. MCL can be used as a global alternative to knowing or estimating each individual IV by IV correlation. One value can be used for a global multicollinearity. The final two parameters, NB and SB, can be used to change and set path values (beta values). NB is the number of specified standardized regression coefficients that the researcher desires to set as known. SB is the actual path values to be entered into the model. These values need to be placed in order in the current macro. In the examples used in this article, the standardized regression coefficients are not set to known values. Some other examples provided in the macro show where this parameter is changed. The outcome of this basic version of the SAS Macro for automating the solution shows the final unknown parameters, calculations for the full correlation matrix, and a simple sample size determination for the standardized regression coefficient that was unknown.

Example 1: Unknown Standardized Partial Regression Coefficients

In the first example, there are no external data and the researcher wants to create a correlation and/or acquire a sample size for a regression model with 4 independent variables, an estimated R^2 value of .5, and an estimated average intercorrelation of $r = 0.3$ Only the first three parameters of the DMATRIX are used, but 0s need to be added to indicate no specified paths are being requested. The PREPDMATRIX is not used. The base matrix %DMATRIX (VN,R2,MCL,FULLMCL,NB,SB) becomes:

%DMATRIX(4, .5, .3, FULLMCL, 0, {0});

The 4 denotes 4 independent variables in the model. The .5 is the R^2 value the researcher chose as an overall estimate. The .3 is the intercorrelation and will be assigned to each IV. FULLMCL is just a place holder and should not be changed here. The first 0 is the number of specified paths the researcher wants to change. The second 0 is the actual path value specified. Effectively, the 0s act as place holders for the parameters as they will not change in this macro example.

The DMATRIX uses the quadratic equation as described in the equations above: $a = 7.6$, $b = 0.0$, and $c = -0.5$. These values are used in the quadratic equation to give a value for all standardized coefficients of 0.25649 in this example. Using Equation 15 (or Equation 8 as all values are equal), we can solve for the correlation of IVs and DV: 0.4873. The sample size for the unknown standardized coefficients in this example is 87.688, which is rounded to 88. This sample size can be seen as the number of subjects needed, given the path value of 0.25649, to have at least one statistically significant path value.

For illustrative purposes, Figure 5 shows that the matrix multiplications can be expressed as described earlier in the quadrant example. Table 1 shows the correlation created from the DMATRIX macro with the standardized coefficients of 0.25649 and correlation of IVs and DV of 0.4873.

Table 1. Full Correlation Matrix.

	X_1	X_2	X_3	X_4	Y
X_1	**1**	**0.30**	**0.30**	**0.30**	**0.49**
X_2	**0.30**	**1**	**0.30**	**0.30**	**0.49**
X_3	**0.30**	**0.30**	**1**	**0.30**	**0.49**
X_4	**0.30**	**0.30**	**0.30**	**1**	**0.49**
Y	**0.49**	**0.49**	**0.49**	**0.49**	**1**

To calculate *a*:

1	1	1	1	1	0.3	0.3	0.3	=	1.9	1.9	1.9	1.9	1	*a* = 7.6
				0.3	1	0.3	0.3						1	
				0.3	0.3	1	0.3						1	
				0.3	0.3	0.3	1						1	

To calculate *b*:

1	1	1	1	0	0	0	0	=	0	0	0	0	1	*b* = 0
				0	0	0	0						1	
				0	0	0	0						1	
				0	0	0	0						1	

To calculate *c*:

1	1	1	1	0	0	0	0	=	0	0	0	0	1	c_1 = 0
				0	0	0	0						1	R^2 0.5
				0	0	0	0						1	$c = c_1 - R^2$
				0	0	0	0						1	*c* = -0.5

Figure 5. Matrix multiplications.

Finally, it would be helpful to see what the model would look like as a standardized path diagram as depicted in Figure 6. Here we see each direct path rounded to .26. The overall R^2 is set to .5 and all the correlations paths are set to .3.

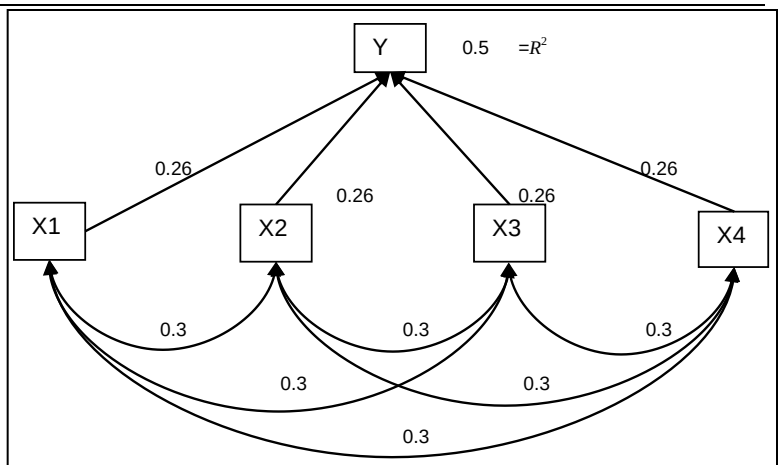


Figure 6. Standardized Path Diagram.

Example 2: One Known Standardized Partial Regression Coefficient

The second example is very similar to the first example. There are no external data and the researcher wants to create a correlation and/or acquire a sample size for a regression model with 4 independent variables and an estimated R^2 value of .5 with a common intercorrelation between IVs of .3. In contrast to example 1, in example 2, the researcher also wants to specify that the first path in the model is .1. This is effectively setting one of the standardized regression coefficients in the model and estimating the remainder.
The base model is changed in a very similar way as in the proceeding example.

%DMATRIX(4, .5, .3, FULLMCL, 1, {.1});

The first 4 parameters are the same as the previous example, but now the first path is specified as .1. In example 1, we set the number of paths to 0 and the value for that path to 0. In this second example, the value 1 indicates one path to be set and .1 is the actual value for that path.

Example 2 shows the manipulation of the matrices for a mixture of known and unknown values. Secondly, it uses the value for which all coefficients would be equal in this model plugged back in as if it were known. As in the previous example, values are used in the quadratic equation: $a = 4.8$, $b = .18$, and $c = -0.49$. The three remaining unknown standardized coefficients are calculated to be 0.301304. The correlation of IVs and DV are 0.3711741 for the known parameter and 0.5120872 for the unknown parameter. The sample size for the unknown standardized coefficients for this example is determined to be 61.845537 or rounded to 62. Note that for the known parameter set to .1 in this example, the sample size would be much higher at 611.

Again, for illustrative purposes, Figure 7 shows that the matrix multiplications can be expressed as described earlier in the quadrant example to see how the model works differently when both known and unknown values are used.

To calculate *a*:

0.1	1	1	1	0	0	0	0	=	0	1.6	1.6	1.6	0.1	*a* = 4.8
				0	1	0.3	0.3						1	
				0	0.3	1	0.3						1	
				0	0.3	0.3	1						1	

To calculate *b*:

0.1	1	1	1	0	0.3	0.3	0.3	=	0.9	0.03	0.03	0.03	0.1	*b* = 0.18
				0.3	0	0	0						1	
				0.3	0	0	0						1	
				0.3	0	0	0						1	

To calculate *c*:

0.1	1	1	1	1	0	0	0	=	0.1	0	0	0	0.1	c_1 = 0.01
				0	0	0	0						*1*	R^2 0.5
				0	0	0	0						*1*	$c = c_1 - R^2$
				0	0	0	0						*1*	*c* = -0.49

Figure 7. Matrix multiplications.

Table 2. Full Correlation Matrix.

	X_1	X_2	X_3	X_4	Y
X_1	**1**	**0.30**	**0.30**	**0.30**	**0.37**
X_2	**0.30**	**1**	**0.30**	**0.30**	**0.51**
X_3	**0.30**	**0.30**	**1**	**0.30**	**0.51**
X_4	**0.30**	**0.30**	**0.30**	**1**	**0.51**
Y	**0.37**	**0.51**	**0.51**	**0.51**	**1**

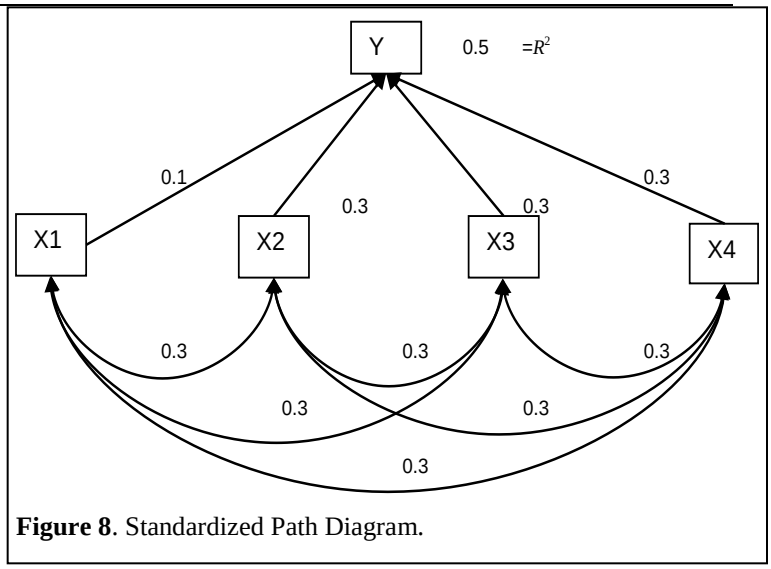


Figure 8. Standardized Path Diagram.

Table 2 shows that the correlation created from the DMATRIX macro with the unknown standardized coefficients of 0.301304. The variables in the correlation matrix associated with a smaller direct path have smaller correlations, while the one path with a larger standardized coefficient has a smaller correlation in this case. In Figure 8, the standardized path diagram is again shown. Here the direct paths that were unknown are rounded to .3 and the known path set by the researcher is .1. The overall R^2 is set to .5 and all the correlations paths are set to .3.

Example 3: Changing the IV by IV Correlations

In this third example, imagine the researcher wants to change everything as an Example 2, but now also wishes to use an external dataset to create the IV by IV matrix instead of setting one global intercorrelation value. This is similar to the first two examples, but instead of simply specifying the .3 value for the MCL parameter, the researcher will use an existing SAS dataset with known values for IV by IV correlations[4]. In this example, we first use PREPDMATRIX. The base macro:

%PREPDMATRIX (IV,GENMCL,FULLMCL);

is used to create the correlation matrix from the raw data. There are 4 IVs in the actual dataset and these need to be renamed IV1, IV2, IV3, IV4. These should be the only variables in the dataset. The dataset is named SASDATASET. The following code can be used to invoke this dataset:

%PREPDMATRIX (4,SASDATASET,FULLMCL);

The outcome of the macro is a correlation matrix called FULLMCL to be used in the DMATRIX. The DMATRIX macro specification would then be:

%DMATRIX(4, .5, 2, FULLMCL, 1, {.1});

This example is similar to Example 2 except notice that the MCL value is set over 1. To use the correlation matrix FULLMCL instead of the MCL value, the MCL needs to be set over 1. In this example, MCL is set to 2 – an impossible value. If the MCL value is set under the absolute value of 1, the MCL parameter will override the FULLMCL parameter. In this example, the matrix of IV correlations comes from a dataset. The quadratic equation values: a = 3.73, b = -0.056, and c = -0.49. The unknown standardized coefficients are calculated to be 0.3699994. The correlation of IVs and DV vary for each variable. The sample size for the unknown standardized coefficients is rounded to 39. To show the full illustrative, Figure 9 simply adds clarification how individual IV by IV correlation values are input to the system.

To calculate *a*:

0.1	1	1	1	0	0	0	0	=	0	0.5	1.87	1.37	0.1	*a* =	3.73
				0	1	0	-0.5						1		
				0	0	1	0.87						1		
				0	-0.5	0.87	1						1		

To calculate *b*:

0.1	1	1	1	0	0.92	-0.4	-0.8	=	-0.28	0.09	-0.04	-0.08	0.1	*b* =	-0.06
				0.92	0	0	0						1		
				-0.4	0	0	0						1		
				-0.8	0	0	0						1		

To calculate *c*:

0.1	1	1	1	1	0	0	0	=	0.1	0	0	0	0.1	c_1 = 0.01
				0	0	0	0						*1*	R^2 0.5
				0	0	0	0						*1*	$c = c_1 - R^2$
				0	0	0	0						*1*	c = -0.49

Figure 9. IV by IV Correlations.

Table 3. Full Correlation Matrix.

	X_1	X_2	X_3	X_4	Y
X_1	**1**	**0.92**	**-0.40**	**-0.80**	**0**
X_2	**0.92**	**1**	**0**	**-0.50**	**0.28**
X_3	**-0.40**	**0**	**1**	**0.87**	**0.65**
X_4	**-0.80**	**-0.50**	**0.87**	**1**	**0.43**
Y	**0**	**0.28**	**0.65**	**0.43**	**1**

Table 3 shows that the correlation created from the DMATRIX macro with the unknown standardized coefficients of 0.3699994. The standardized path diagram is again shown in Figure 10. Here the direct paths that were unknown are rounded to .37 and the known path set by the researcher is .1. The overall R^2 is set to .5 and all the correlations paths have their own value.

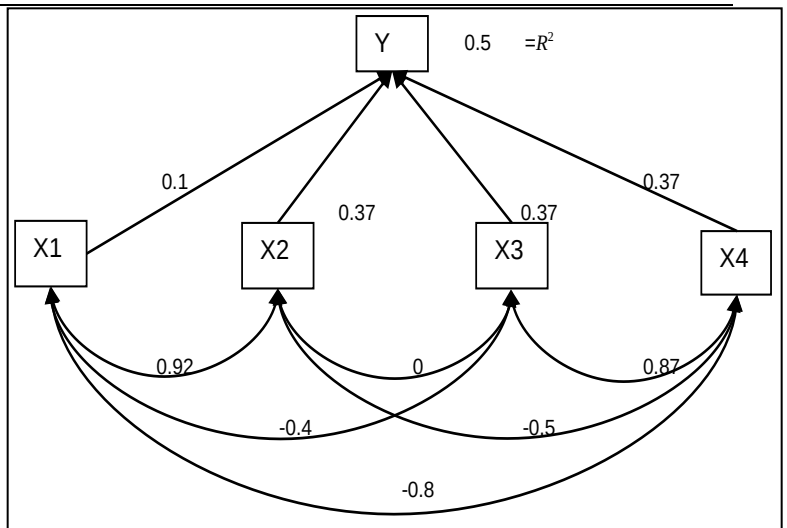


Figure 10. Standardized Path Diagram.

Another advantage to the matrix formulation is that it is very easy to use any structure for the matrix of IV correlations. The only adaptation that is required is simply setting the appropriate quadrants of the IVs correlation matrix to zero and known IV correlations to any value. The IVs intercorrelation matrix can essentially be entirely known *a priori* or estimated separately as desired. The SAS macros make using this method in a relatively straightforward manner.

Additional Considerations and Discussion

In this study, we have developed an automated path tracing program for general linear models that allows the calculation of unknown parameters given an assumption of equality for unknown beta

coefficients. The assumption of coefficient equality is not intended to be interpreted as a naïve expectation that beta values will be equal in applied research, but rather as a reasonable approach when parameters are unknown. The article outlines the linear and matrix algebra equations by which automation of path analysis can occur. The quadratic formula was used to calculate the positive root (the negative root can also be calculated) of a beta coefficient given the known parameters and assumptions of the modeling process.

This automation provides a tool for both the methodological researcher and the applied practitioner. The methodological researcher, in conducting Monte Carlo simulations, can use the instrument to develop a full correlation matrix from estimation of as little as two known parameters. Building a correlation matrix in this manner, notwithstanding the limitations of equality of parameters, can facilitate the data generation process at the heart of many methodological studies such as the R^2 shrinkage formulae study conducted by Fan et al. (2001). The parameters used in their example: R^2, multicollinearity values, and number of predictors could be investigated at many levels in a straightforward way. For the applied researcher, this automation provides a method to reasonably estimate parameters in a model and the sample size needed to acquire these parameters. These methodological and applied investigations can be extended to other statistics of the GLM to include analysis of variance and covariate models.

Data generation is but one application for this type of path tracing algorithm. Another beneficial by-product of the entire structure is the ability to derive a sample size necessary to find an effect of a particular size under pre-specified Type I error rate (α) and level of achieved power (1-β).In the case of parameter equality – all unknown coefficients are set to be equal – this value can be used as an expected effect size. Furthermore, if R^2 were true, this effect should yield a maximum practical sample size. Essentially, in practice, we know our parameters will not be equal even when we have a reasonable approximation of the overall R^2 value. This value of equality is, however, the smallest any individual parameter can be (assuming they are all set as unknown) to still achieve a desired overall effect in the model. In our previous example, we had expected coefficients of size 0.2565. This information can be utilized in subsequent sample size calculations. For example, suppose $\alpha = 0.05$, the desired level of power is 0.80 (*Z*-adjusted 0.84), and the use of a one tailed *Z*-test value = 1.645 (one could be more stringent by using a *t*-value, maximizing at 1*df*, but here the overall *Z*-adjusted is 0.84+1.645=2.485). Our effect size in this case can be the value in the example – 0.2565 with $R^2 = 0.5$. The sample size can be calculated as:

$$Sample\ size = N = \frac{(2.485)^2(1-0.2565^2)}{(0.2565)^2} \approx 88 \text{ (rounded up)}$$

A sample size can be computed for each of the four paths given the standardized regression coefficients, desired R^2 value, the specified level of power, and Type I error control. This type of computation could be looked at as the maximum sample size desired to find an effect on any of the paths if the overall effect, or better, is reached in the real world application. Of course, sample size is subjected to change upon any modification of the other elements in its calculation (i.e., overall effect R^2, increasing or decreasing the standardized regression coefficient, etc.).

The basic methodological approach presented here can be extended in a natural way to multivariate analysis and structural equation modeling. While this approach has straightforward applicability for GLMs, its value in applied multivariate settings will increase in areas where estimates of sample size and power are more difficult to compute, and where overall model development is more complicated for the researcher to create. To this end, methodological work in the general multivariate setting is already underway with the intent of demonstrating how these extensions can also benefit from automated path tracing rules established in this article.

Endnotes

1. Appendices and SAS macro accompanying the manuscript can be obtained from the author's website: http://education.umd.edu/EDMS/fac/Harring/webpage.html#publications
2. Again, the SAS macro is available from the second author's website.
3. The notation used here for R^2, β, and r is slightly different than that used in the article, but corresponds to the same entities.
4. The SAS dataset used can also be found at the second author's website: http://education.umd.edu/EDMS/fac/Harring/webpage.html#publications

References

Cohen, J., Cohen, P., West, S. G., & Aiken, L. S. (2003). *Applied multiple regression/correlation analysis for the behavioral sciences* (3rd ed.). Mahwah, NJ: Erlbaum.

Fan, X., Felsővályi, À., Sivo, S. A., & Keenan, S. C. (2001). *SAS® for Monte Carlo studies: A guide for quantitative researchers*. Cary, NC: SAS Institute Inc.

Harwell, M. R., Stone, C. A., Hsu, T.-C., & Kirisci, L. (1996). Monte Carlo studies in item response theory. *Applied Psychological Measurement, 20,* 101-125.

Hutchinson, S. R., & Bandelos, D. L. (1997). A guide to Monte Carlo simulation research for applied researchers. *Journal of Vocational Education Research, 22,* 233-245.

Loehlin, J. C. (1998). *Latent variable model: An introduction to factor, path, and structural analysis* (3rd ed.). Mahwah, NJ: Erlbaum.

Pedhazur, E. J., & Pedhazur-Schmelkin, L. (1991). *Measurement, design, and analysis: An integrated approach*. Mahwah, NJ: Erlbaum.

Wright, S. (1960). Path coefficients and path regression: Alternative or complementary concepts? *Biometrics, 16,* 189-202.

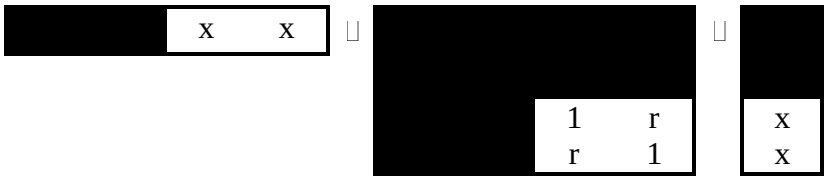
Send correspondence to: Jeffrey R. Harring
University of Maryland
Email: harring@umd.edu

APPENDIX

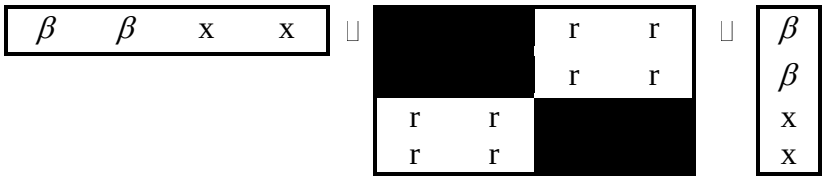
Matrix Expressions

In the following expressions, β's are known parameters, x's are unknown parameters, and r's are correlations among the IVs. The coefficients of the quadratic equation in matrix notation are specified as:

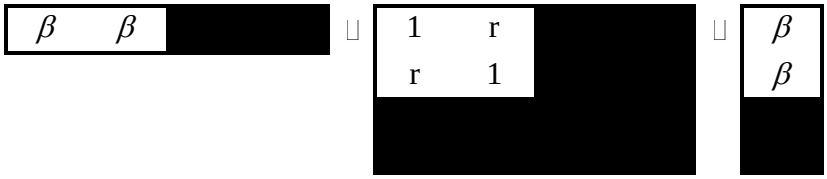
$a = \boldsymbol{\beta}_u' \mathbf{R}_{Q_4} \boldsymbol{\beta}_u$



$b = \boldsymbol{\beta}_p' \mathbf{R}_{Q_2 Q_3} \boldsymbol{\beta}_p$



$c = \boldsymbol{\beta}_k' \mathbf{R}_{Q_1} \boldsymbol{\beta}_k - R^2$



R^2 must be subtracted from this matrix expression.