Automated Path Tracing for General Linear Models

Objectives

The proposed research has three main purposes: (1) to algebraically derive a kernel for calculating direct causal effects in a path model for the univariate general linear model, specifically discussing regression models and extensions such as ANOVA, (2) to provide, in this context, a rationale for decisions made for dealing with multiple unknown parameters and multicollinearity, (3) to demonstrate, through empirical examples, how these derived tools can be utilized to help generate data under a myriad of realistic conditions as well as to provide needed information for sample size/power calculations. Extensions to other classes of models (e.g., generalized linear models) will be discussed.

Theoretical Framework

Monte Carlo simulation is an invaluable and versatile tool in those methodological situations where analytic techniques are inadequate. 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, \( \text{Var}(y) \) is known. However, when \( \text{Var}(y) \) is unknown, the impact of substituting an estimate of \( \text{Var}(y) \) 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 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, 1997), there are
numerous other potentially beneficial outcomes that could be explored, and are often tailored to reflect nuances of a particular methodology (i.e., structural equation modeling, item response theory). Finally, the literature seems replete with studies which 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 $\rho^2$. 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 to specify these correlation values to obtain the correct overall $R^2$, and perhaps more importantly for multiple regression simulations, how should regression coefficients be specified so that the model can be properly specified. 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.

**Methods**

Path Tracing for Regression Models

Multiple regression requires a basic understanding of sample statistics ($n$, 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

$$
\hat{z}_{y_j} = \beta_1 z_{x_{i1}} + \beta_2 z_{x_{i2}} + \cdots + \beta_p z_{x_{ip}}
$$

$$
\hat{y}_j = \sum_{k=1}^{p} \beta_k z_{x_{ik}}
$$

(1)
where $\beta_k$ denotes the $k$th standardized regression coefficient.

In the language of path analysis, standardized partial regression coefficient 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.

The illustration in Figure 1 is a simple path diagram of a regression model with 2 independent variables, $X$, and one dependent variable $Y$. Using Wright’s rules (Wright, 1960) or through manipulation of beta equations (Loehlin, 1998), the above path diagram can be solved for the coefficient of determination or $R^2$

$$R^2 = b_1^2 + b_2^2 + 2b_1b_2r_{12}$$

(2)

where, $R^2$ is the overall coefficient of determination value for a multiple regression equation. In this case the multiple regression has two independent variables (IV’s) and one dependent variable (DV). The standardized partial regression coefficients or path coefficients are $b_1$ and $b_2$, and $r_{12}$ is the correlation between the two IV’s $b_1$ and $b_2$. Extending this same model to 3 IV’s results in

$$R^2 = b_1^2 + b_2^2 + b_3^2 + 2b_1b_2r_{12} + 2b_1b_3r_{13} + 2b_2b_3r_{23}$$

(3)

where $b_1$, $b_2$, and $b_3$ are the path coefficients, $r_{12}$, $r_{13}$, and $r_{23}$ are the correlations between the IV’s. While the expression for $R^2$ becomes progressively more convoluted as the number of
IV’s increases, a relatively straightforward pattern begins to emerge. A generalized pattern for an expression of $R^2$ with $p$ IV’s is

$$R^2 = \sum_{j=1}^{p} b_j^2 + 2 \sum_{j=1}^{p} \sum_{k=1}^{p} b_j b_k r_{jk}.$$  \hspace{1cm} (4)

Two kernels of patterns are present in the generalized expression for $R^2$ in (4). The squared path coefficients, $\nu R^2 = b_1^2 + b_2^2 + \ldots + b_p^2 = \sum_{j=1}^{p} b_j^2$, can be interpreted as the variance component of the model where $\nu R^2$ is the portion of the coefficient of determination derived from the variance components of the model, and where $j$ indexes the coefficient corresponding to the $j$th IV. The portion of the coefficient of determination attributed to covariance defines the second kernel: $c R^2 = 2b_1 b_2 r_{12} + 2b_1 b_3 r_{13} + \ldots + 2b_k b_j r_{kj} = 2 \sum_{j=1}^{p} \sum_{k=1}^{p} b_j b_k r_{jk}$.

Traditionally, $R^2$ is written as

$$R^2 = \sum_{j=1}^{p} r_{yx} \beta_j$$  \hspace{1cm} (5)

where $r_{yx}$ represents the correlation between the DV and $j$th IV, and $\beta_j$ is the standardized beta coefficient. One could rewrite equation (5) without the DV and IV correlations and instead use only the correlations amongst the IV’s

$$R^2 = \sum_{j=1}^{p} \sum_{k=1}^{p} \beta_j r_{jk} \beta_k$$  \hspace{1cm} (6)

where $\beta_j$ and $\beta_k$ are the standardized beta coefficient and $r_{jk}$ is the correlation between two IV’s variables $x_j$ and $x_k$ over $j = 1, \ldots, p$ and $k = 1, \ldots, p$ for all $j$ and $k$.

**Solving for Beta**

In order to solve an arbitrary regression equation with multiple unknown beta coefficients a theoretical decision needs to be made in 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, \( \beta's \), are equal equations (3) and (5) simplify to

\[
R^2 = \sum_{j=1}^{p} \beta_j^2 + 2 \sum_{j=1}^{p} \sum_{k=1}^{p} \beta_j \beta_k r_{jk} \tag{7}
\]

\[
R^2 = \sum_{j=1}^{p} \sum_{k=1}^{p} \beta_j r_{jk} \beta_k \tag{8}
\]

respectively. Each beta value corresponds to one IV. As an example, imagine the case with \( p = 4 \) and all of the variances set equal to one another. The expression for \( R^2 \) can be written as

\[
R^2 = b_1^2 + b_2^2 + b_3^2 + b_4^2 + 2b_1b_2r_{12} + 2b_1b_3r_{13} + 2b_1b_4r_{14} + 2b_2b_3r_{23} + 2b_2b_4r_{24} + 2b_3b_4r_{34}.
\]

The variance portion of the model simplifies to

\[
R^2 = 4b^2 + 2b_1b_2r_{12} + 2b_1b_3r_{13} + 2b_1b_4r_{14} + 2b_2b_3r_{23} + 2b_2b_4r_{24} + 2b_3b_4r_{34}.
\]

Further assuming all correlations amongst the IV’s are the same yields

\[
R^2 = 4b^2 + 2r(b_1b_2 + b_1b_3 + b_1b_4 + b_2b_3 + b_2b_4 + b_3b_4)
\]

and because all path coefficients \( b \) are equal, the expression for \( R^2 \) simplifies to

\[
R^2 = 4b^2 + 2r(6(b^2))
\]

\[
= 4b^2 + 12rb^2
\]

\[
= b^2(4 + 12r).
\]

For a specific value for \( R^2 \), the path coefficient is the positive root of the quadratic

\[
b^2 = \frac{R^2}{4 + 12r}, \text{ namely}
\]

\[
b = \sqrt{\frac{R^2}{4 + 12r}}.
\]

While the result above is specific for a regression model with 4 predictors, solving for \( \beta \) can be easily generalized for any number of predictors given the restrictions that the coefficients are identical and the correlations between the IV’s are also the same.

\[
R^2 = \beta^2 p + \beta^2 r [p(p - 1)] \tag{9}
\]
Solving the expression for $R^2$ in (9) for $\beta$ results in

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

(10)

where $\beta$ 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**

Using the formula in (10) the correlations from the full set of coefficients can be constructed as

$$r_{jk} = b + b(p - 1)r$$

(11)

where $r$ is the same overall correlation between the independent variables, and $r_{jk}$ is the correlation between DV and the $k$th IV for each IV related to the DV. All correlations will be the same and is a null condition of no difference. To see how this system is interconnected, suppose we wanted to solve for the $b$’s in the path diagram shown in Figure 2 with overall $R^2 = 0.5$ and $r = 0.3$.

*Figure 2. Path diagram relating one dependent variable to two independent variables with $R^2 = 0.5$ and correlation between the independent variables, $r = 0.3$*
Using (10)

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

This can be checked with the expression in (1):

\[ R^2 = b_1^2 + b_2^2 + 2b_1b_2r_{12} = R^2 = 0.438^2 + 0.438^2 + 2 \cdot 0.438 \cdot 0.438 \cdot 0.3 = 0.5 \]

The correlation between IV and DV can be computed use equation:

\[ r_{yk} = b + b(p-1)r = 0.438 + 0.438 \cdot (2-1) \cdot 0.3 = 0.570 \]

In the context of simulating data to examine bias correction formulae for regression sample \( R^2 \), eight IV’s are used to construct a data correlation matrix.

\[ \beta = \sqrt{\frac{R^2}{p + r[p(p-1)]}} = b = \frac{.5}{8 + .3\left[\frac{8(8-1)}{8-1}\right]} = 0.142 \]

Again, this value can be confirmed with (1):

\[ R^2 = \beta^2 p + \beta^2 r[p(p-1)] \]
\[ = 0.142^2 \cdot 8 + 0.142^2 \cdot 0.3\left[\frac{8(8-1)}{8-1}\right] \]
\[ = 0.5 \]

The correlation between the DV and any IV is computed as:

\[ r_{yk} = b + b(p-1)r \]
\[ = 0.142 + 0.142 \cdot (8-1) \cdot 0.3 \]
\[ = 0.4402 \]

These were the same values obtained in a demonstration of how to execute a Monte Carlo simulation by Fan, Felsovalyi, Sivo, and Keenan (2002). In contrast to the ad-hoc, guess-and-check method apparently employed by Fan et al., the values obtained above can be constructed in a straightforward manner.
**Preliminary Results**

Regression models are comprised, among other things, structural elements such as correlations and coefficients. In the context of data generation, the two previous examples highlighted how these interconnected pieces can be solved for algebraically to adhere to a particular underlying structure (see equations (9) – (11)). These linear equations have been worked out in SAS Interactive Matrix Language (IML) as matrix algebra expressions. A SAS macro has been developed that can solve for paths more robustly than expressed in the linear equations (which will be provided in the full paper’s Appendix). The matrix equations, which will be fully developed in the paper, are operationalized in the SAS programming and permit separate correlations between predictors allowing existing datasets or correlation matrices a researcher may have available to be used. In addition, path coefficients can be set. Two very useful methodological applications of these equations have been set up in the Macro: (1) to solve for power, and (2) to develop a matrix such as the one used in the multiple regression example by Fan et al.

**Scientific Importance**

Monte Carlo simulation has proven to be an effective tool for investigating methodological problems in those situations where analytic solutions are impractical or simply do not exist. An integral ingredient of a simulation design is generating data under specific model configurations and/or distributional assumptions. For general linear models, this often means setting values for overall fit, correlations between variables, and regression coefficients. To our knowledge, no one as demonstrated in a systematic fashion how to specify these values to follow a particular structure. Showing how this can be accomplished algebraically certainly makes this process more transparent. In addition, having tools, like the proposed SAS macro that automates this process, will be beneficial for methodologists seeking to investigate statistical properties associated with the general linear model.
References


