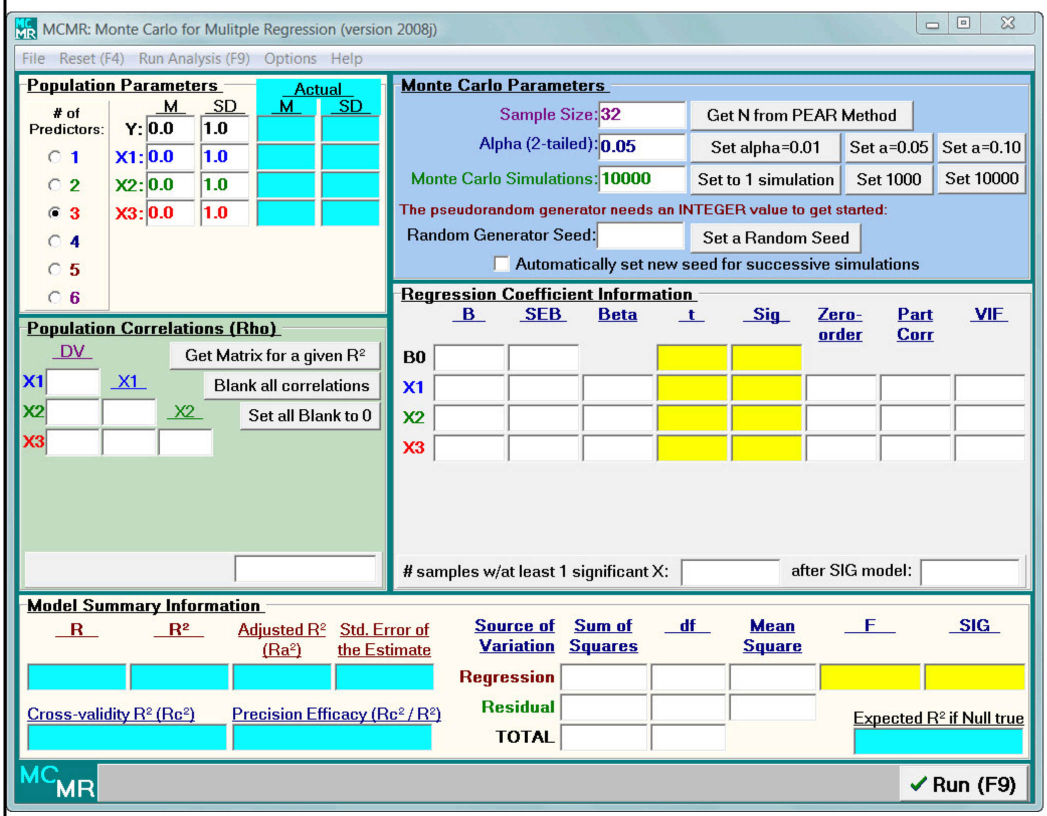
A Monte Carlo Program for Multiple Linear Regression

Gordon P. Brooks
Ohio University

The primary purpose of this presentation is to demonstrate a new computer program that statistics instructors can use to help teach certain regression topics in their courses. In particular, a computer program was written in Borland Delphi 2007 and will run under most recent versions of the Microsoft Windows operating system, including XP and Vista. The program may be downloaded free of charge.

The MCMR: Monte Carlo for Multiple Regression program performs Monte Carlo simulations of ordinary least squares multiple linear regression with up to 6 predictors. The program runs single sample analyses in addition to Monte Carlo simulations. For single samples, data can be saved and imported in comma-delimited text format. For Monte Carlo analyses, sampling distribution data can be saved for several regression statistics for further analyses elsewhere. The on-screen results from any analysis can be saved to a file and printed. The summary results provided from the Monte Carlo simulations include R-squared statistics, shrinkage statistics, regression coefficients, standard errors, and other relevant statistical results. Suggestions for use will be provided to help users understand how the program can be used effectively in intermediate statistics courses.

The MCMR Program is available at: http://oak.cats.ohiou.edu/~brooksg/software.htm

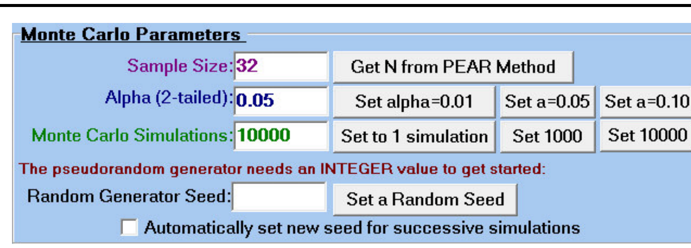


This is the Opening Screen that appears when the program is started (or after the "Reset (F4)" menu option is chosen).

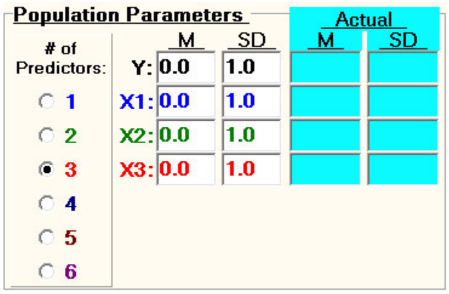
3 sections require user input

This is where we describe the population from which samples will be drawn in the Monte Carlo process. That is, the Monte Carlo process randomly generates samples of data that could come from the particular population described (using means, standard deviations, and correlations).

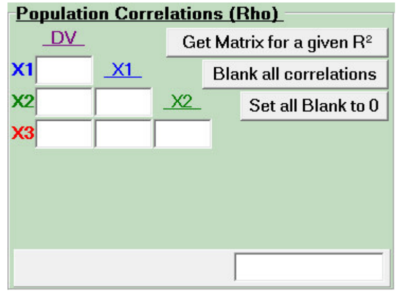
Click "Run" (bottom right) or press F9 to begin the Monte Carlo analysis.



Set sample size, alpha, number of simulations, and maybe a seed for the random number generator (if you use the same seed, you get the same results).

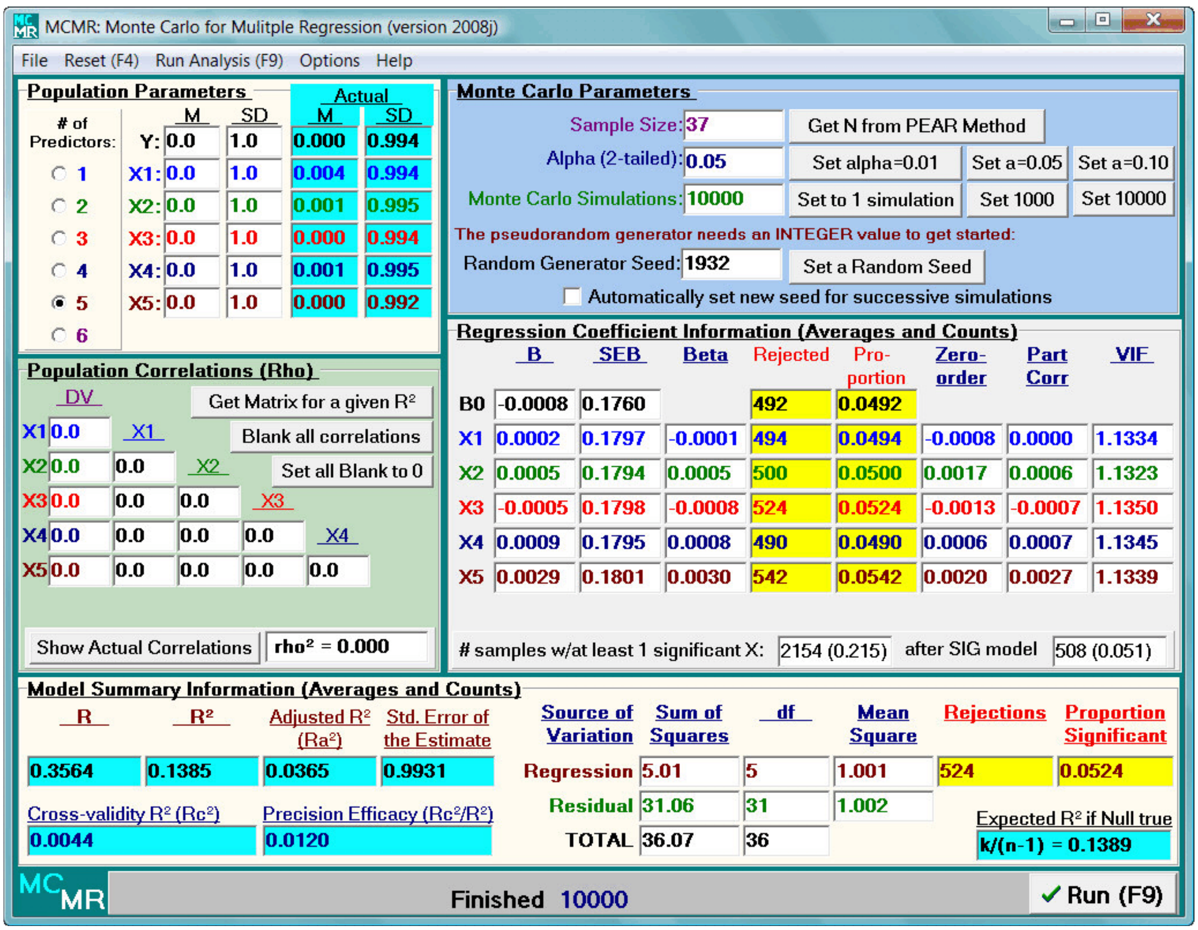


Choose the number of predictors and Set the population means and standard deviations (Y is the dependent variable, X1 is predictor 1, etc.)



Set the population correlations (rho). You can get a random matrix that meets certain criteria (described later). Some matrices will not work as proper CORRELATION MATRICES. If one is entered, and error message will pop up, saying that the matrix is not Positive Definite (see Get Matrix section below).

After an analysis



4 boxes contain results after an analysis, but not all are immediately obvious. Each section is described in greater detail below. This analysis was done with a seed of 1932. All population correlations were 0.0.

Population Parameters

# of Predictors:		M	SD	Actual M	Actual SD
	Y:	0.0	1.0	0.000	0.994
○ 1	X1:	0.0	1.0	0.004	0.994
○ 2	X2:	0.0	1.0	0.001	0.995
○ 3	X3:	0.0	1.0	0.000	0.994
○ 4	X4:	0.0	1.0	0.001	0.995
◉ 5	X5:	0.0	1.0	0.000	0.992
○ 6					

The average ACTUAL means and standard deviations are reported in aqua.

AVERAGE Sample Correlations

	DV	X1	X2	X3	X4
X1	-0.001				
X2	0.002	0.001			
X3	-0.001	-0.001	0.000		
X4	0.001	0.002	0.000	-0.002	
X5	0.002	0.002	-0.001	0.002	0.004

Show Pop. Correlations | rho² = 0.000

If you hit the “Show Actual Correlations” button, you can see the average ACTUAL correlations. (*You must hit “Show Pop. Correlations” to run another analysis*)

Regression Coefficient Information (Averages and Counts)

	B	SEB	Beta	Rejected	Proportion	Zero-order	Part Corr	VIF
B0	-0.0008	0.1760		492	0.0492			
X1	0.0002	0.1797	-0.0001	494	0.0494	-0.0008	0.0000	1.1334
X2	0.0005	0.1794	0.0005	500	0.0500	0.0017	0.0006	1.1323
X3	-0.0005	0.1798	-0.0008	524	0.0524	-0.0013	-0.0007	1.1350
X4	0.0009	0.1795	0.0008	490	0.0490	0.0006	0.0007	1.1345
X5	0.0029	0.1801	0.0030	542	0.0542	0.0020	0.0027	1.1339

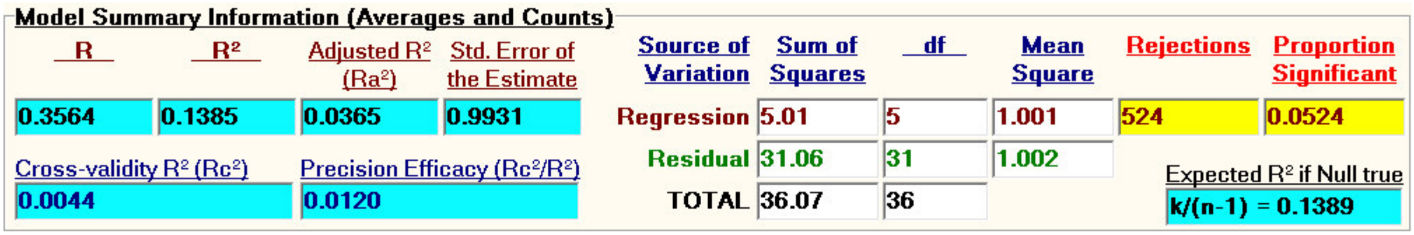
\# samples w/at least 1 significant X: 2154 (0.215) after SIG model 508 (0.051)

The average ACTUAL regression coefficient information is reported in this box — except for the “Rejected” and “Proportion” columns, which report the number (and proportion) of samples in which the particular regression coefficient (represented by X1, X2, etc.) was statistically significant.

“# samples w/at least 1 significant X” reports how many samples had at least one significant predictor.

“after SIG model” reports how many samples had at least one significant predictor following a significant overall regression model (the idea being that we don’t usually examine the statistical significance of regression coefficients unless the model was first significant—but that doesn’t mean that some predictors weren’t significant anyway).

B0 represents the CONSTANT in the regression equation. By default, B0 is not included in the 2 counts (above), but there is a menu option that will allow it to be included.



Model Summary Information (Averages and Counts)

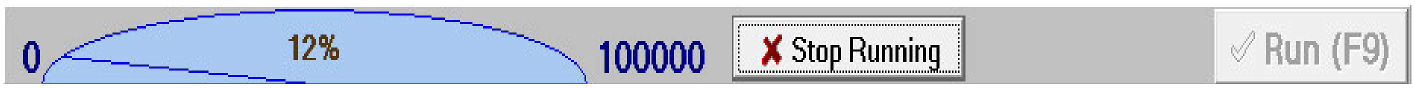
R	R²	Adjusted R² (Ra²)	Std. Error of the Estimate
0.3564	0.1385	0.0365	0.9931

Cross-validity R² (Rc²)	Precision Efficacy (Rc²/R²)
0.0044	0.0120

Source of Variation	Sum of Squares	df	Mean Square	Rejections	Proportion Significant
Regression	5.01	5	1.001	524	0.0524
Residual	31.06	31	1.002		
TOTAL	36.07	36			

Expected R² if Null true: k/(n-1) = 0.1389

Model summary information is provided here. Again, these are AVERAGE results except for the “Rejections” and “Proportion Significant” columns, which report how many (and the proportion of) samples that had statistically significant overall regression models.

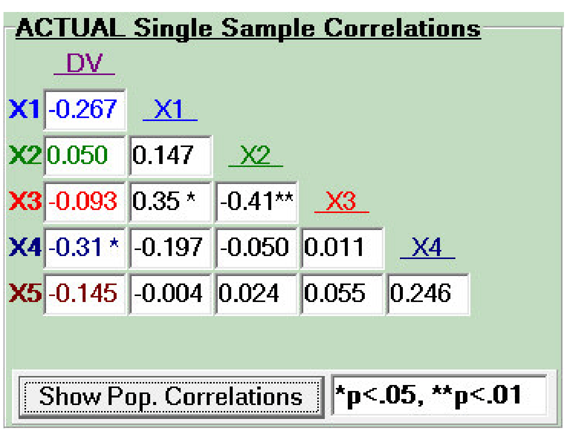


While the Monte Carlo simulations are running, the bottom panel (progress bar) looks like this. You can stop the Monte Carlo analyses if you need to by clicking the “Stop Running” button.

After the analysis is finished, the bottom panel will look like this. If you have aborted the process by pressing the “Stop Running” button, the number actually finished will appear in the panel.

If you review the ACTUAL correlations by clicking on the “Show Actual Correlations” button, you will not be able to continue with additional Monte Carlo analyses until you press the “Show Pop. Correlations” button (which is actually the same button as the “Show Actual Correlations” button.

Although not done in this example, when you run multiple SINGLE SAMPLE analyses, you will have the option of going backwards by one sample. Often, you get to clicking the “Run” button too quickly and you aren’t able to stop on a sample with interesting results. This “Back Up” button will allow you to go back 1 sample (but only 1).



ACTUAL Single Sample Correlations

	DV	X1	X2	X3	X4
X1	-0.267				
X2	0.050	0.147			
X3	-0.093	0.35 *	-0.41**		
X4	-0.31 *	-0.197	-0.050	0.011	
X5	-0.145	-0.004	0.024	0.055	0.246

Show Pop. Correlations   *p<.05, **p<.01

Another difference for SINGLE SAMPLE analyses is that statistically significant pairwise correlations are marked with asterisks when you click “Show Actual Correlations.”

Regression Coefficient Information (Single Sample)

	B	SEB	Beta	t	Sig	Zero-order	Part Corr	VIF
B0	0.0300	0.1699		0.1766	0.8610			
X1	-0.3296	0.1523	-0.3974	-2.1649	0.0382	-0.2665	-0.3422	1.3480
X2	0.1560	0.2108	0.1366	0.7400	0.4649	0.0499	0.1170	1.3634
X3	0.1075	0.1918	0.1097	0.5607	0.5790	-0.0928	0.0886	1.5318
X4	-0.3338	0.1524	-0.3659	-2.1908	0.0361	-0.3094	-0.3463	1.1159
X5	-0.0686	0.1708	-0.0658	-0.4017	0.6907	-0.1449	-0.0635	1.0721

At least 1 significant predictor (X) ? YES after SIG model? No

For SINGLE SAMPLE analyses, the "Rejected" and "Proportion" columns change to the actual *t* statistics and *p* values ("Sig") for each regression coefficient.

By the way, "B" is the unstandardized regression coefficient, "SEB" is the standard error for the unstandardized regression coefficient, "Beta" is the standardized regression coefficient, "Zero-order" is the Pearson correlation between each predictor and Y, "Part Corr" is the part (or semi-partial) correlation between each predictor and Y GIVEN the other predictors in the model, and "VIF" is the variance inflation factor (1/Tolerance) used for diagnosing multicollinearity.

The "At least 1 significant predictor (X) ?" box shows whether any of the regression coefficients was statistically significant (but not which one).

Both bottom boxes turn from white to GREEN if "YES"

Model Summary Information (Single Sample)

R	R²	Adjusted R² (Ra²)	Std. Error of the Estimate
0.4746	0.2253	0.1003	0.9766

Cross-validity R² (Rc²)	Precision Efficacy (Rc²/R²)
0.0000	0.0000

Source of Variation	Sum of Squares	df	Mean Square	F	Sig
Regression	8.60	5	1.719	1.8029	0.1414
Residual	29.57	31	0.954		
TOTAL	38.16	36			

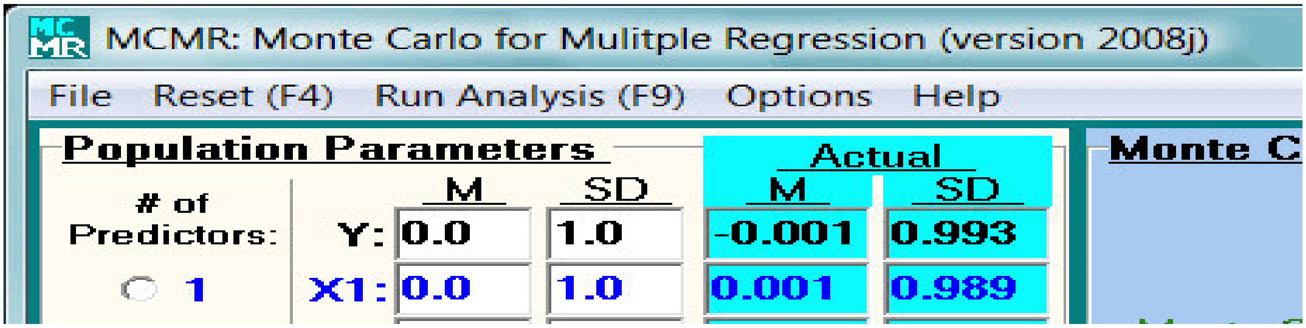
Expected R² if Null true: k/(n-1) = 0.1389

For SINGLE SAMPLE analyses, the "Rejections" and "Proportion Significant" columns change to the actual *F* statistic and the actual *p* value significance of the regression model ("Sig").

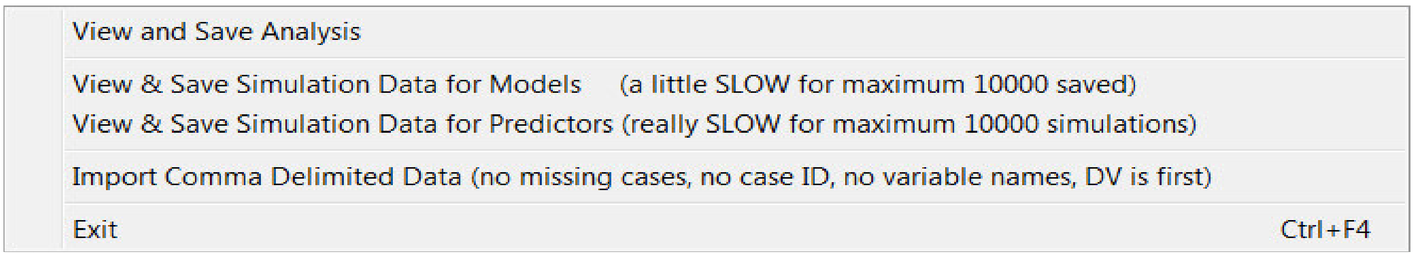
If the model is statistically significant, the "F" and "Sig" boxes turn from yellow to GREEN. If Adjusted R^2 or Cross-validity R^2 are negative they are set to 0.0 (theoretically, neither they nor R^2 can be negative).

By the way, the "Expected R^2 if Null True" box uses the calculation presented by Herzberg (1969), $k/(n-1)$, to show the bias of the R^2 statistic. The "Options" menu allows you to change the information reported here to a few other things.

Menus



"File," "Options," and "Help" show sub-menus (below), but "Reset (F4)" and "Run Analysis (F9)" just perform the given action. "Reset (F4)" will return the program to the main opening screen and "Run Analysis (F9)" will run the analysis, just like clicking the "Run (F9)" button or pressing the F9 key.



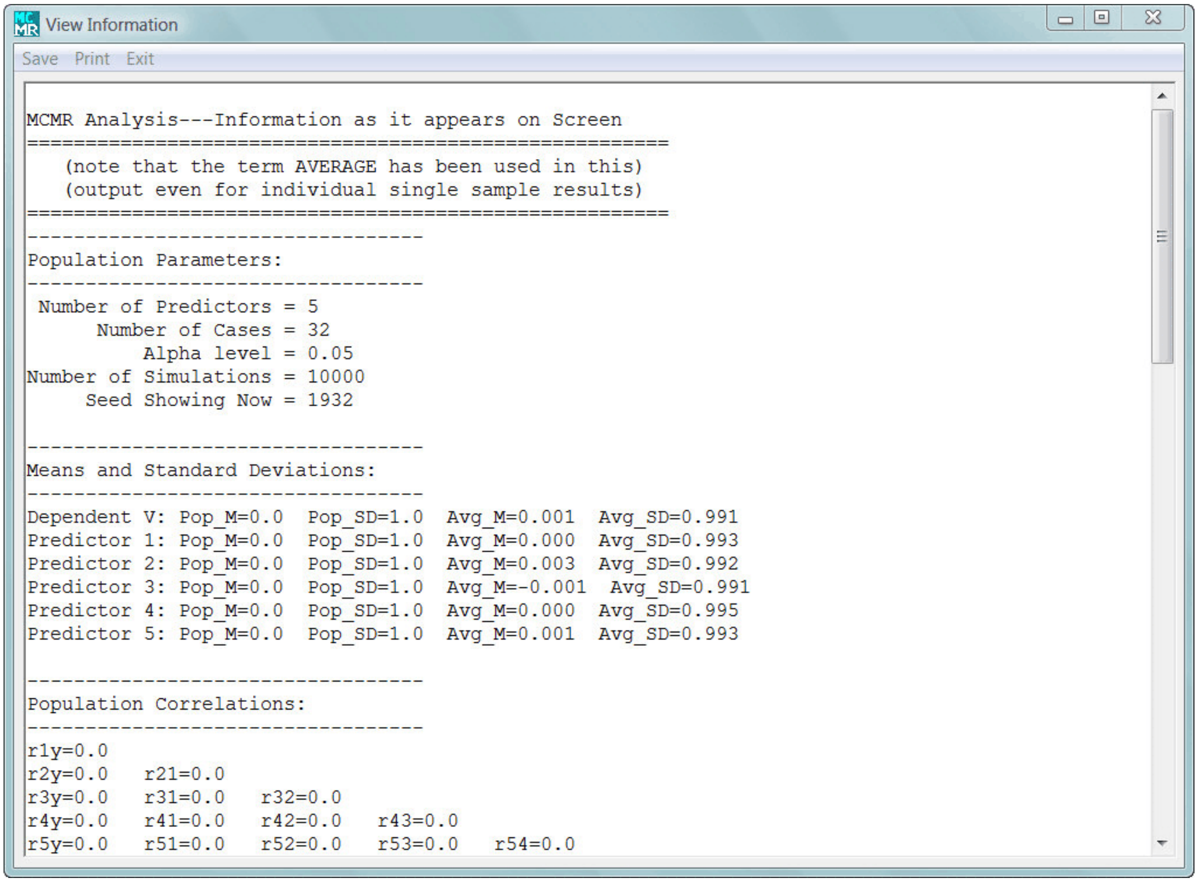
"View and Save Analysis" will show a text version of the results in another window (below), which will also allow you to save and print the results of the analysis.

"View & Save Simulation Data for Models" will save the Model Summary statistics (e.g., R^2, Standard Error of the Estimate) from all the Monte Carlo simulated samples (up to a maximum of 10,000) for analysis in any program that accepts Comma-Delimited text files. Variable names ARE included on the first line of the file.

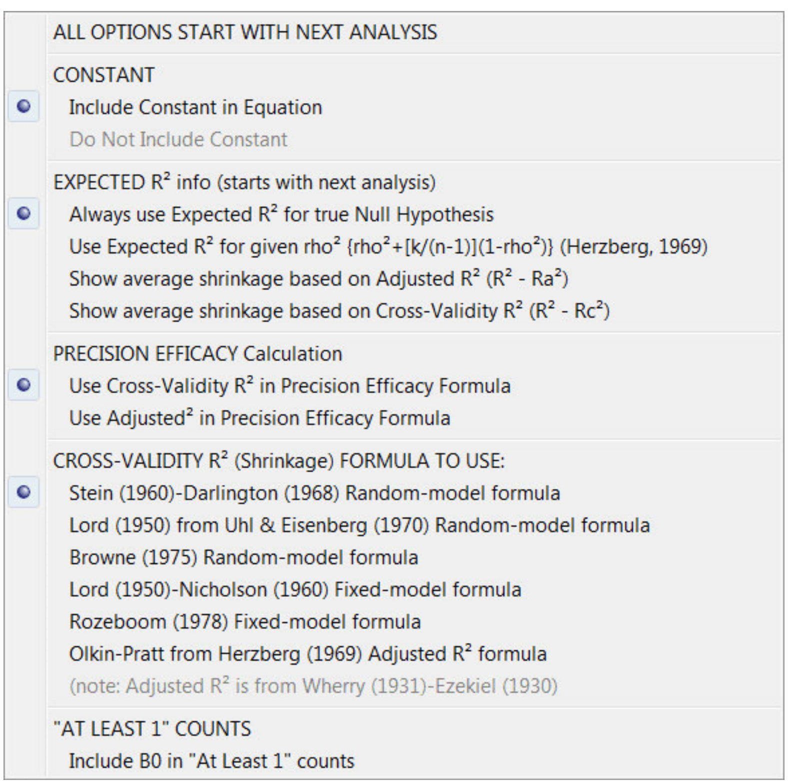
"View & Save Simulation Data for Predictors" will save the Regression Coefficient statistics (e.g., B, SEB, Beta) from all the Monte Carlo simulated samples (up to a maximum of 10,000) for analysis in any program that accepts Comma-Delimited text files. Variable names ARE included on the first line of the file.

If you are running a SINGLE SAMPLE analysis, there is also an option to save SINGLE SAMPLE data. The data from the current single sample analysis is saved WITHOUT variable names on the first line.

"Import Comma Delimited Data" will allow you to read in data that you have previously saved with MCMR, or will allow you to import data saved in appropriate format from any other program (e.g., a spreadsheet or statistics program). The MCMR program assumes that NO variable names are listed on the first line—that is, the data begin on line 1.



All "View and Save" options will open this window. From here, you can "Save" or "Print" the information in the window (using the appropriate menu option).

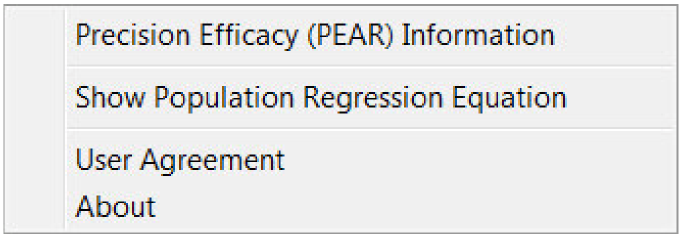


Currently, only analyses with the Constant Included in the Equation are permitted.

There are 4 types of information that can be reported in the box that by default is labeled "Expected R^2 if Null true" — 2 for expected R^2 and 2 for shrinkage.

Precision Efficacy (Brooks, 1998) is calculated using Cross-Validity R^2 by default, but could be calculated using Adjusted R^2. (see help menu for additional information about Precision Efficacy) Different formulas can be used to calculate Cross-Validity R^2 — 6 are available here.

You can choose to have significant B0 included in the counts reported (by default it is not).

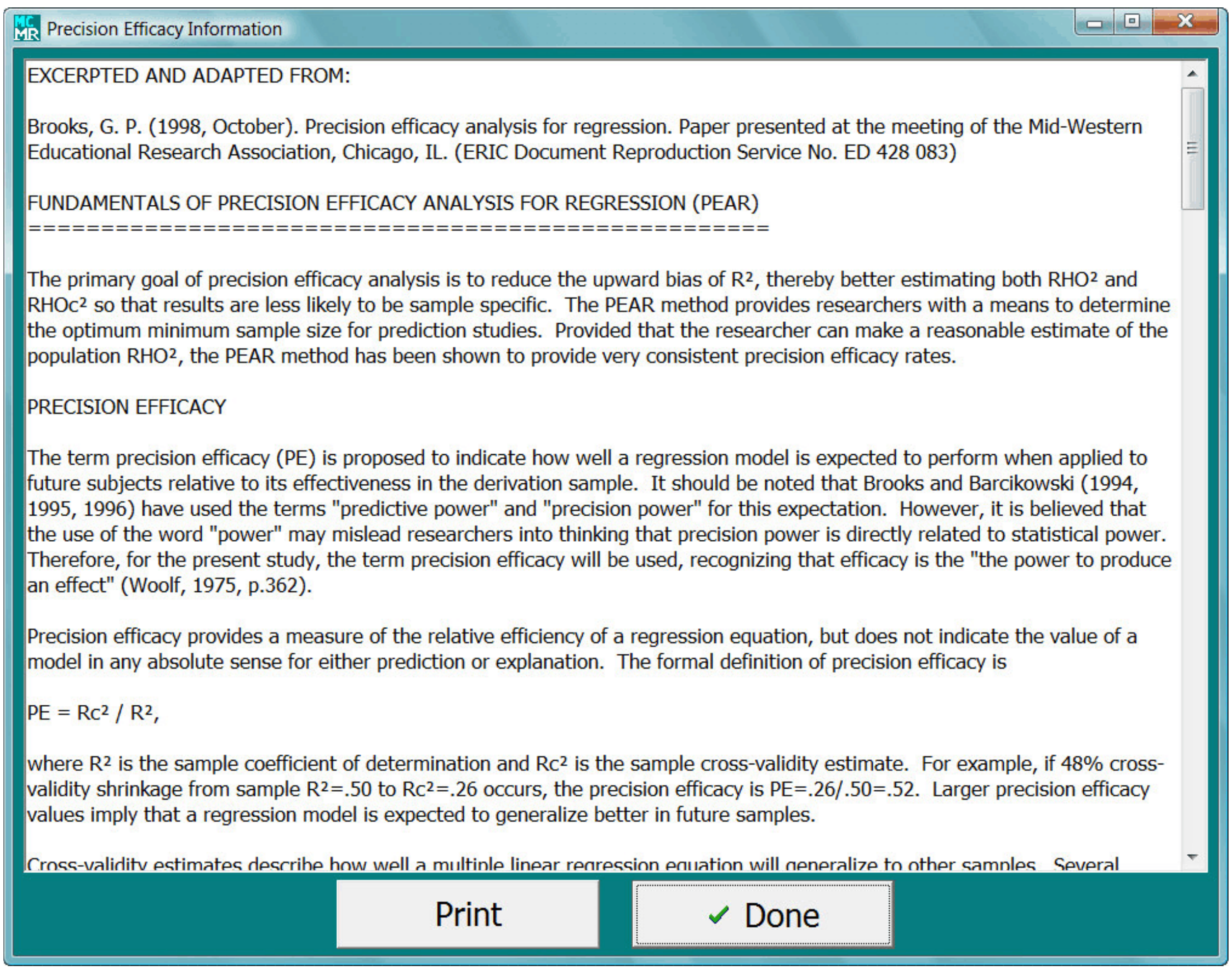


The “Precision Efficacy (PEAR) Information” option will open a window that contains an except from a paper written in 1998 (see below).

“Show Population Regression Equation” will show the STANDARDIZED regression model based on the Population Correlation matrix used to generate data for the analysis.

“User Agreement” opens a window with LICENSE information (important).

“About” provides some basic information about the MCMR program.



EXCERPTED AND ADAPTED FROM:

Brooks, G. P. (1998, October). Precision efficacy analysis for regression. Paper presented at the meeting of the Mid-Western Educational Research Association, Chicago, IL. (ERIC Document Reproduction Service No. ED 428 083)

FUNDAMENTALS OF PRECISION EFFICACY ANALYSIS FOR REGRESSION (PEAR)
===

The primary goal of precision efficacy analysis is to reduce the upward bias of R^2, thereby better estimating both RHO^2 and $RHOc^2$ so that results are less likely to be sample specific. The PEAR method provides researchers with a means to determine the optimum minimum sample size for prediction studies. Provided that the researcher can make a reasonable estimate of the population RHO^2, the PEAR method has been shown to provide very consistent precision efficacy rates.

PRECISION EFFICACY

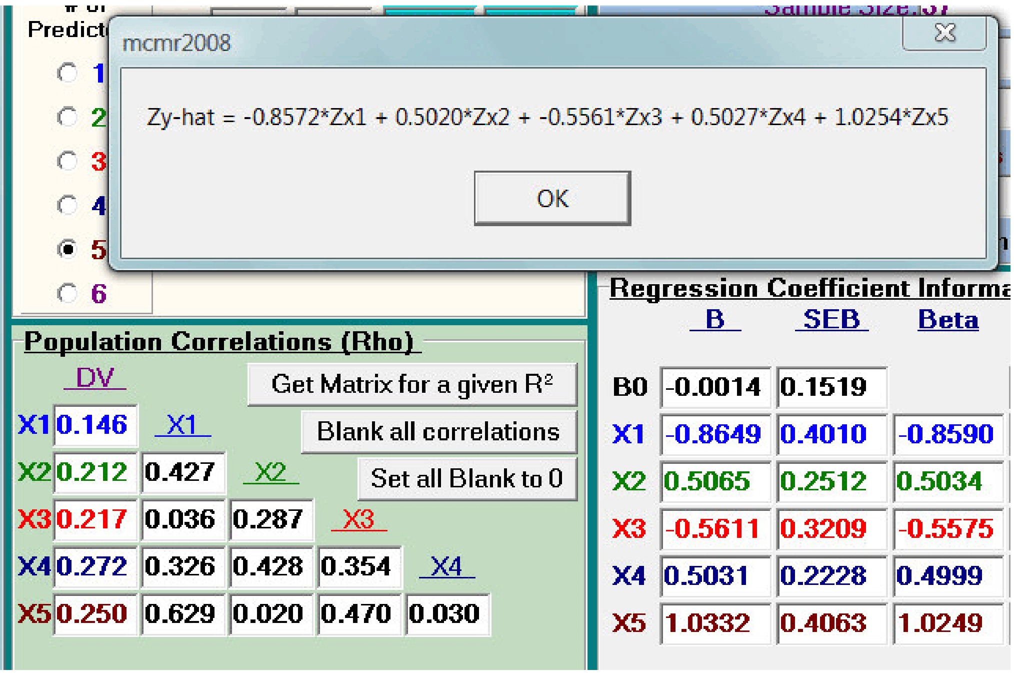
The term precision efficacy (PE) is proposed to indicate how well a regression model is expected to perform when applied to future subjects relative to its effectiveness in the derivation sample. It should be noted that Brooks and Barcikowski (1994, 1995, 1996) have used the terms "predictive power" and "precision power" for this expectation. However, it is believed that the use of the word "power" may mislead researchers into thinking that precision power is directly related to statistical power. Therefore, for the present study, the term precision efficacy will be used, recognizing that efficacy is the "the power to produce an effect" (Woolf, 1975, p.362).

Precision efficacy provides a measure of the relative efficiency of a regression equation, but does not indicate the value of a model in any absolute sense for either prediction or explanation. The formal definition of precision efficacy is

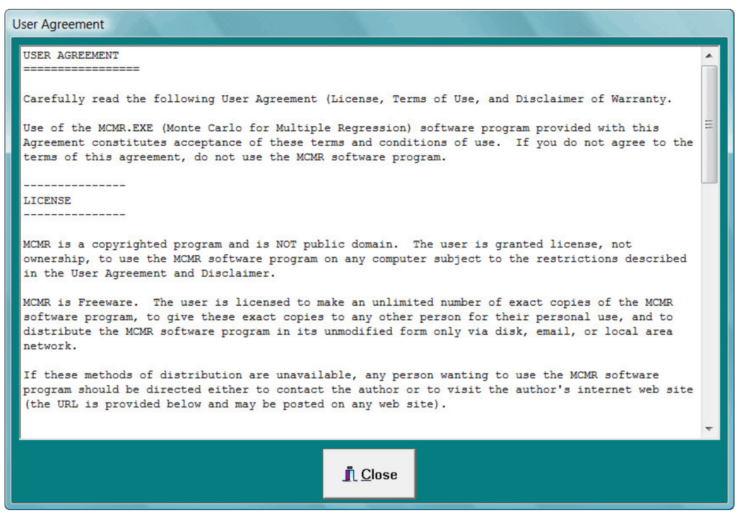
$PE = Rc^2 / R^2$,

where R^2 is the sample coefficient of determination and Rc^2 is the sample cross-validity estimate. For example, if 48% cross-validity shrinkage from sample R^2=.50 to Rc^2=.26 occurs, the precision efficacy is PE=.26/.50=.52. Larger precision efficacy values imply that a regression model is expected to generalize better in future samples.

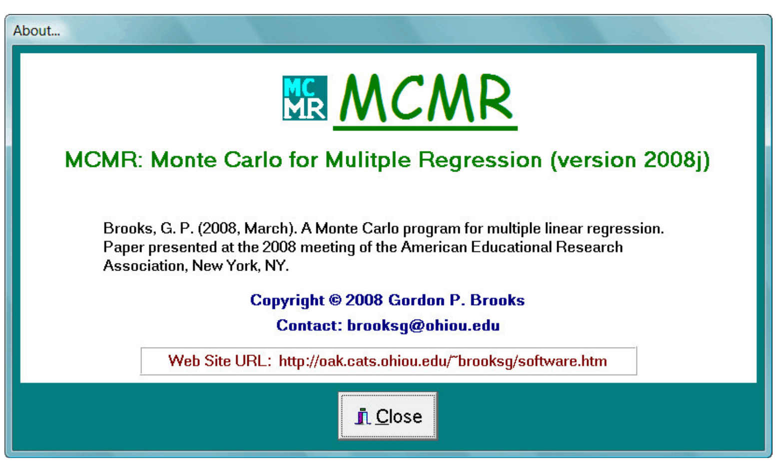
Cross-validity estimates describe how well a multiple linear regression equation will generalize to other samples. Several



At any point, the user can request this pop-up window that shows the Population Standardized Regression model for comparison to current results.

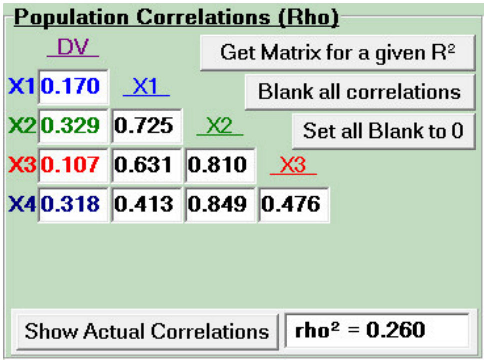


The User Agreement window, with important information about the legal use of the software.

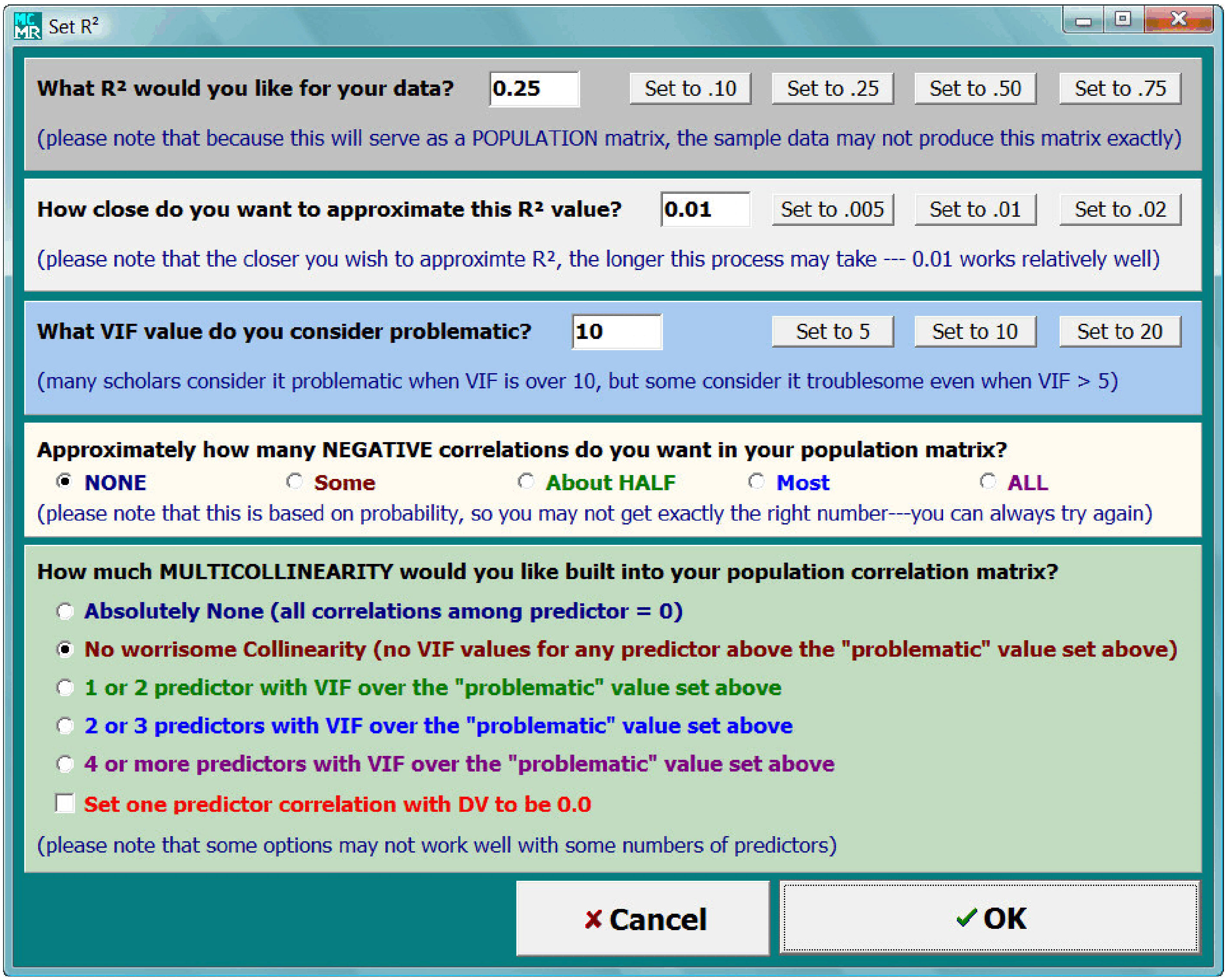


The ABOUT window with the full name of the program, copyright and contact information, and the web site from which this and other software programs may be obtained.

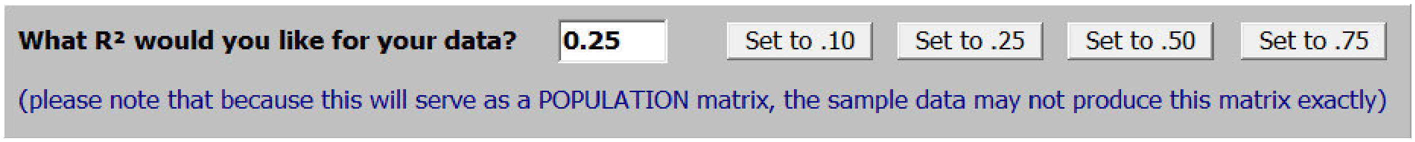
Secondary Window: Get a Population Matrix with certain Given Characteristics



If you click the "Get Matrix for a Given R2" button, the following window will open — allowing you to get a correlation matrix that meets certain criteria.

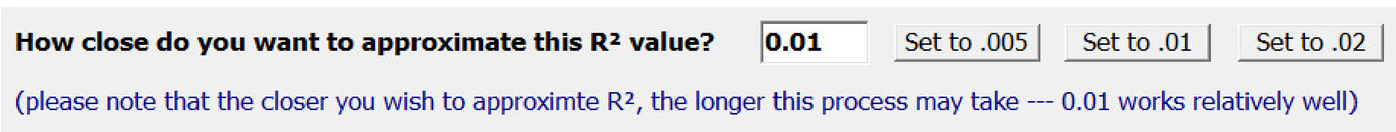


Each section is described more below. When you click "OK" a correlation matrix will be found with the given criteria (if possible) AND that correlation matrix will be transferred to the main MCMR program screen into the "Population Correlations (rho)" section.



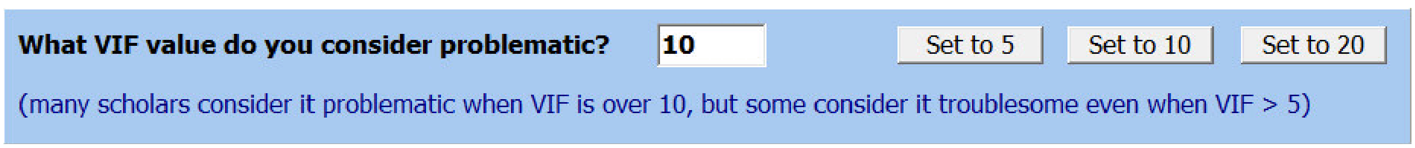
You can choose any R^2 for your POPULATION correlation matrix (so really this is a rho^2 or ρ^2 value), but buttons are provided for some common values (these are based on tables from Park & Dudycha, 1974).

Remember, however, that this will derive a POPULATION correlation matrix, from which samples will be drawn during the Monte Carlo process. This value says nothing specific about any of the R^2 values calculated in the samples (other than they should be from the population with the derived population correlation matrix).



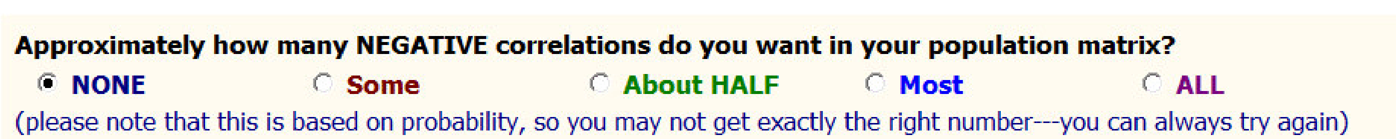
You can choose how close you want to approximate the population R^2 set in the previous box. While it is indeed possible to approximate some matrices very closely, anything smaller than 0.001 will likely take a good deal of time. The values 0.01, or even 0.005, seem to work pretty well if you really want to get exact.

Remember, however, that this is how closely you approximate the desired population R^2 in the POPULATION correlation matrix, and says nothing about the samples drawn during the Monte Carlo process.

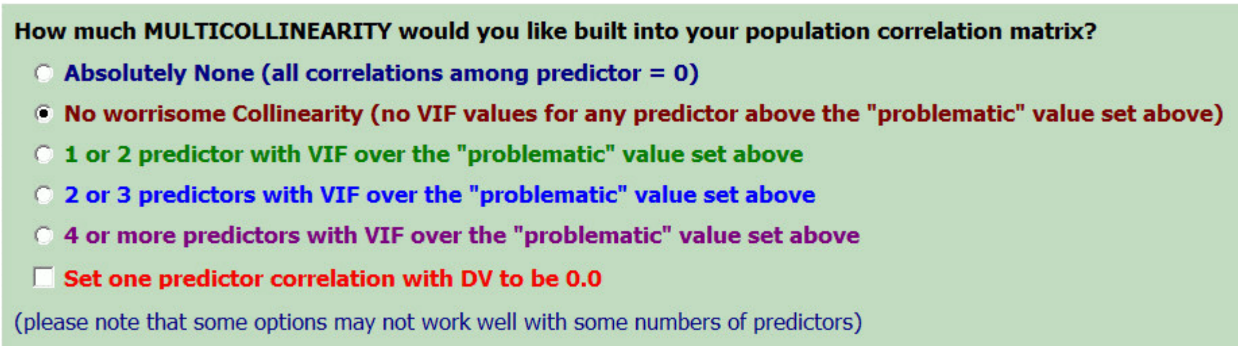


You can set any value above 1.0 for the critical VIF threshold value. Most scholars choose 5.0 or 10.0, depending on how much MULTICOLLINEARITY(also called COLLINEARITY) you're willing to tolerate.

Recall that **VIF = 1/Tolerance**, where **Tolerance** $= 1 - \mathbf{R}_j^2$, where $\mathbf{R}_j^2$ is the squared correlation when the $\mathbf{j}^{th}$ predictor acts as a temporary dependent variable being predicted by all the other predictors.



This option will allow you to create a population correlation matrix with some (or many) negative correlations.



This box will allow you to request a certain level of multicollinearity in your population correlation matrix.

$ “Absolutely None” requires that all intercorrelations among predictors are 0.0, but the correlations between the predictors and Y will be set randomly to provide the R^2 given above.

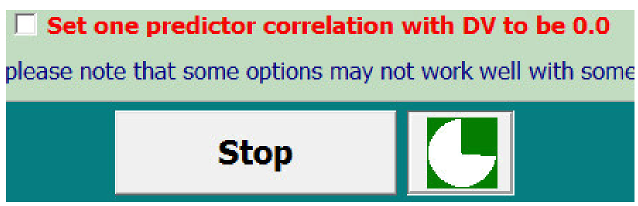
$ “No Worrisome Collinearity” will produce a population correlation matrix where all predictor intercorrelations will be non-zero, but will be probably smaller than the critical VIF set above.

$ “1 or 2 predictors with VIF” will produce a population correlation matrix such that predictor intercorrelations will probably result in at least 1, but not more than 2, VIF values over the critical value

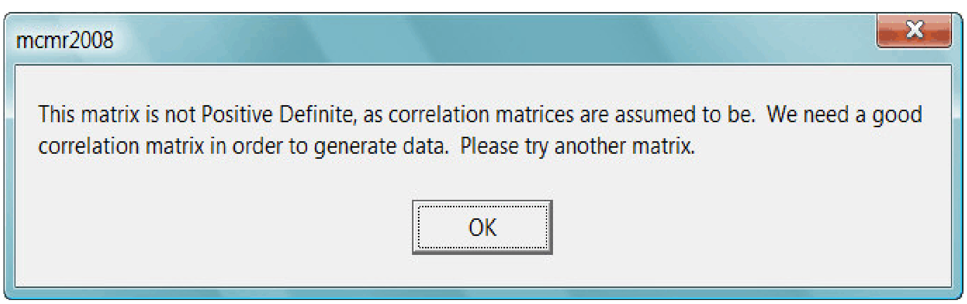
$ “2 or 3 predictors with VIF” will produce a population correlation matrix such that predictor intercorrelations will probably result in at least 2, but not more than 3, VIF values over the critical value

$ “4 or more predictors with VIF” will produce a population correlation matrix such that predictor intercorrelations will probably result in at least 4 VIF values over the critical value

Note that “probably” was included in these descriptions. There are rare occasions, given certain starting correlations used in the algorithm, where the resulting correlation matrix does not match the criteria exactly. You can either go ahead and use the derived matrix, or simply try another. Different seeds used in each run of this sub-program result in different matrices being created.

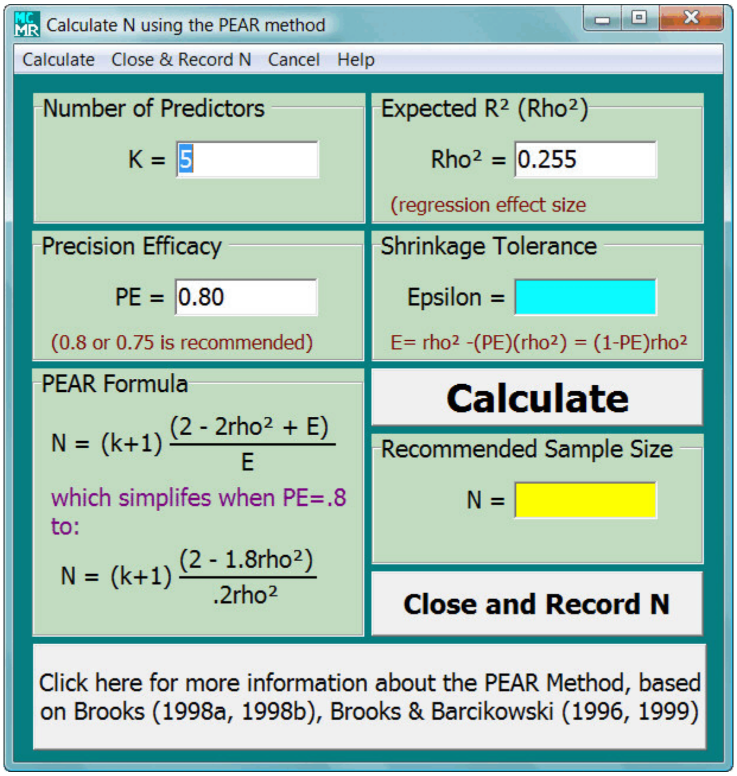


If the little pie ever fills in all GREEN during this process, you probably have a matrix that cannot be created. You can try a few more times, if you’d like, because sometimes different seeds do produce workable results. You can also allow the program to continue running for a while, which sometimes will produce a workable result (the algorithm continues to adjust itself a little as it runs, which sometimes allows results to work).



This error message will be shown whenever the “Stop” button is pushed (above), whenever the user has entered an inappropriate matrix, or on very rare occasions where rounding the derived correlations to 3 decimal places impacts the matrix enough to make it unusable.

Secondary Window: Get a Sample Size using the PEAR Method



The user can change the parameters of the PEAR method (Brooks, 1998). By default, this window will provide the information for the analysis in the main window, if possible. For example, once the number of predictors is determined, it will be filled in here. Note that any number of predictors can be inserted.

More information about Precision Efficacy (PE) and the Precision Efficacy Analysis for Regression (PEAR) sample size method can be found by clicking the "Click here for more information" button (see below).

Briefly, however, Precision Efficacy is a complement to Proportional Shrinkage based on an appropriate Cross-validity R^2 (R_C^2) formula. Shrinkage itself (ε, or epsilon) can be written as

$$\boldsymbol{\varepsilon = R^2 - R_C^2}$$

whereas Proportional Shrinkage (PS) might be written as $\mathbf{PS = (R^2 - R_C^2) / R^2}$

Precision Efficacy would therefore be PE = 1 - PS, or

$$\mathbf{PE = R_C^2 / R^2}$$

Solving PE = 1 – ε/R² for ε, and replacing R² with an expected, a priori Re², results in the formula

$$\boldsymbol{\varepsilon = R_E^2 - (PE)(R_E^2)}$$

where R_E^2 is often just set at the expected population ρ^2. Because Precision Efficacy (PE) is usually set at .75 or .80, shrinkage would usually be $.25\rho^2$ or $.2\rho^2$, respectively. Note that shrinkage may also be set absolutely as something like ε = .05 or ε = .10.

Once parameters are set, "Calculate" will determine the required sample size. The recommended sample size will appear in the YELLOW box underneath the "Calculate" button.

"Close and Record N" will move this sample size to the main screen.

"Cancel" (on the menu bar) will close the dialog window without making any changes to the main screen.

Although the PEAR method was derived using Cross-Validity R^2 (Brooks, 1998), it is theoretically reasonable to apply the same idea to Precision Efficacy calculated using Adjusted R^2 instead. Algina and Olejnik (2000) have discussed a similar idea, but different approach, to sample sizes for Adjusted R^2.

In this case, sample sizes would be determined such that the SHRINKAGE from R^2 to Adjusted R^2 would be maintained within a certain range. For example, if R^2 is .25, then Adjusted R^2 would be at least .20 when Precision Efficacy of .80 was used as the criterion. The formula for sample sizes to be used with such an approach would be

$$\mathbf{N = (k + 1)(1 - R_E^2 + \varepsilon) / \varepsilon}$$

Where

R_E^2 = expected population ρ^2
k = number of predictors
$\varepsilon = (R^2 - R_A^2)$

as compared to

$$\mathbf{N = (k + 1)(2 - 2R_E^2 + \varepsilon) / \varepsilon}$$

where **$\varepsilon = (R^2 - R_C^2)$**

for Cross-Validity (see Brooks, 1998). Shrinkage tolerance can also be calculated as

$$\mathbf{\varepsilon = (1 - PE)\, R^2}$$

where, for PE = .80, it would simplify to (just like it would also for the Cross-Validity approach)

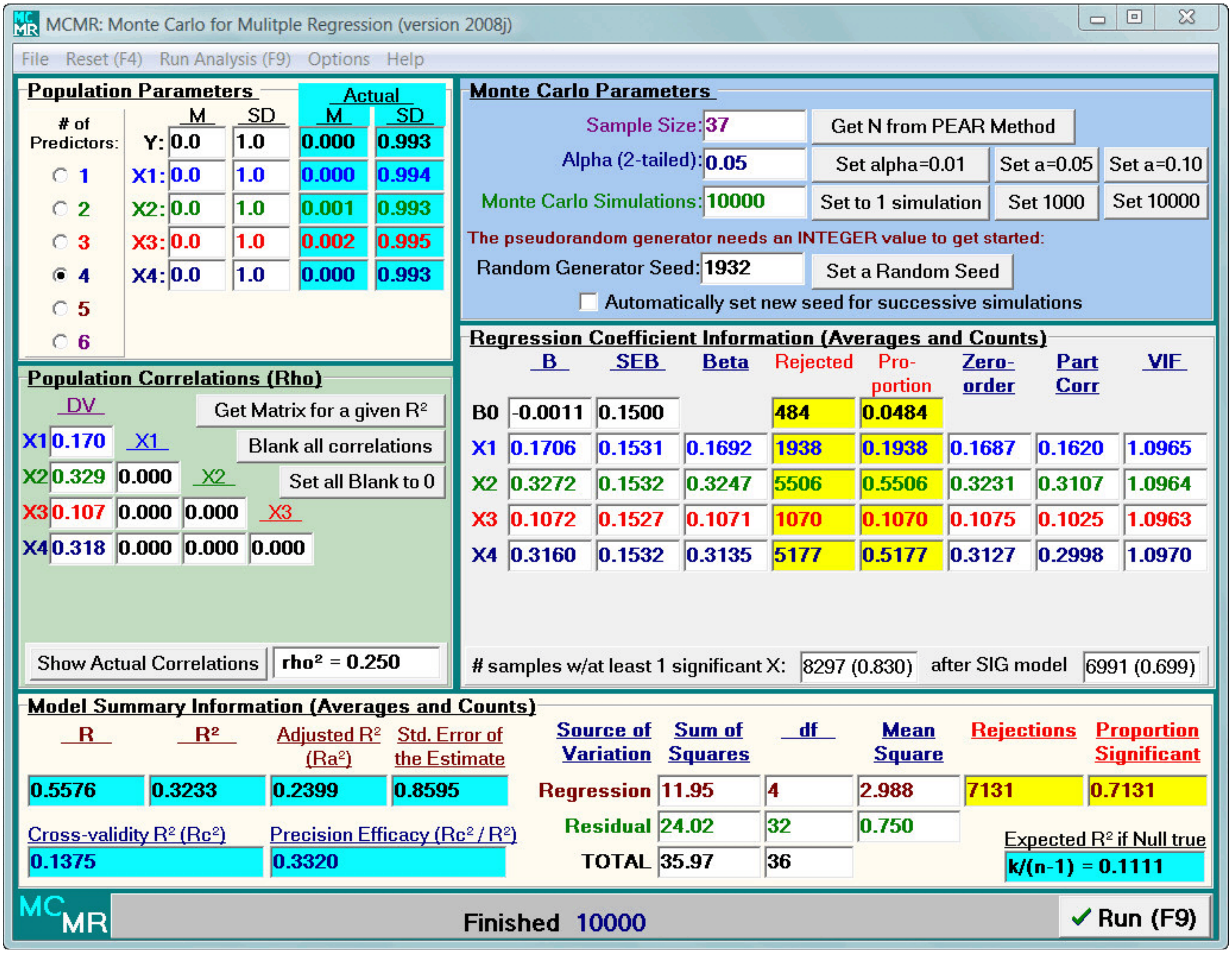
$$\mathbf{\varepsilon = .2R^2}$$

Recall that one of the options on the “Options” menu is to use Adjusted R^2 in the Precision Efficacy formula instead of Cross-Validity R^2.

The key difference is that for Cross-validity Precision Efficacy, the idea is to INCREASE Cross-validity R^2; however, for Adjusted R^2, the idea is more to DECREASE R^2, making it closer to the true population parameter (since Adjusted R^2 is usually a good estimate of rho^2).

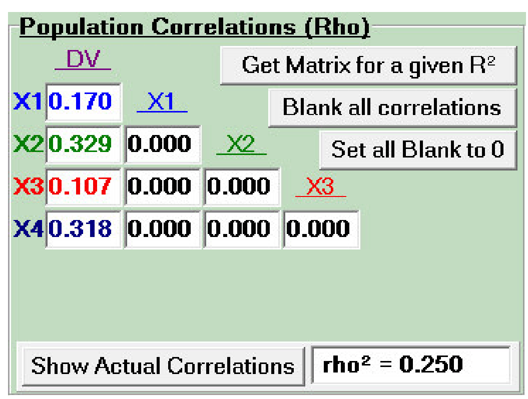
Either method helps make the model more generalizable by decreasing the standard errors for the regression coefficients. The Corss-validity approach is more stringent because it accounts for error not only in the regression model derivation sample, but also for the error in future samples to which the regression model is applied.

An Example: Multicollinearity and Inflation of Standard Errors



Let's assume that all predictor intercorrelations are 0.0, while predictor correlations with Y are non-zero such that rho^2 = .25.

The seed is set to 1932, with N = 37, alpha = .05, and 10,000 simulated samples are drawn.



Regression Coefficient Information (Averages and Counts)

	B	SEB	Beta	Rejected	Proportion	Zero-order	Part Corr	VIF
B0	-0.0011	0.1500		484	0.0484			
X1	0.1706	0.1531	0.1692	1938	0.1938	0.1687	0.1620	1.0965
X2	0.3272	0.1532	0.3247	5506	0.5506	0.3231	0.3107	1.0964
X3	0.1072	0.1527	0.1071	1070	0.1070	0.1075	0.1025	1.0963
X4	0.3160	0.1532	0.3135	5177	0.5177	0.3127	0.2998	1.0970

samples w/at least 1 significant X: 8297 (0.830) after SIG model 6991 (0.699)

In this case, the standard errors for the regression coefficients ("SEB") are each approximately 0.153. Note that the Variance Inflation Factors ("VIF") are all roughly 1.096—since there is no correlation among the predictors we would expect this to be near 1.0, but since each of the 10,000 samples drawn probably ad some minor correlation among the predictors, it will not be exactly 1.0.

If we arbitrarily add some correlation among the predictors, BUT LEAVE THE CORRELATIONS BETWEEN THE PREDICTORS AND Y THE SAME, we introduce multicollinearity.

Note that in this matrix, the ρ^2 is not exactly .250 any more. This is arbitrary, but will have some minor impact on our results.

In particular, if you examine the model summary results (we won't here), you would see some minor differences — especially in R^2 and the Sum of Squares due to the regression (which impacts other things as well). This is not a REAL difference, but rather due to the different population conditions set by the slightly larger ρ^2.

Population Correlations (Rho)

	DV	X1	X2	X3
X1	0.170			
X2	0.329	0.725		
X3	0.107	0.631	0.810	
X4	0.318	0.413	0.849	0.476

Get Matrix for a given R²

Blank all correlations

Set all Blank to 0

Show Actual Correlations

$\rho^2 = 0.260$

Regression Coefficient Information (Averages and Counts)

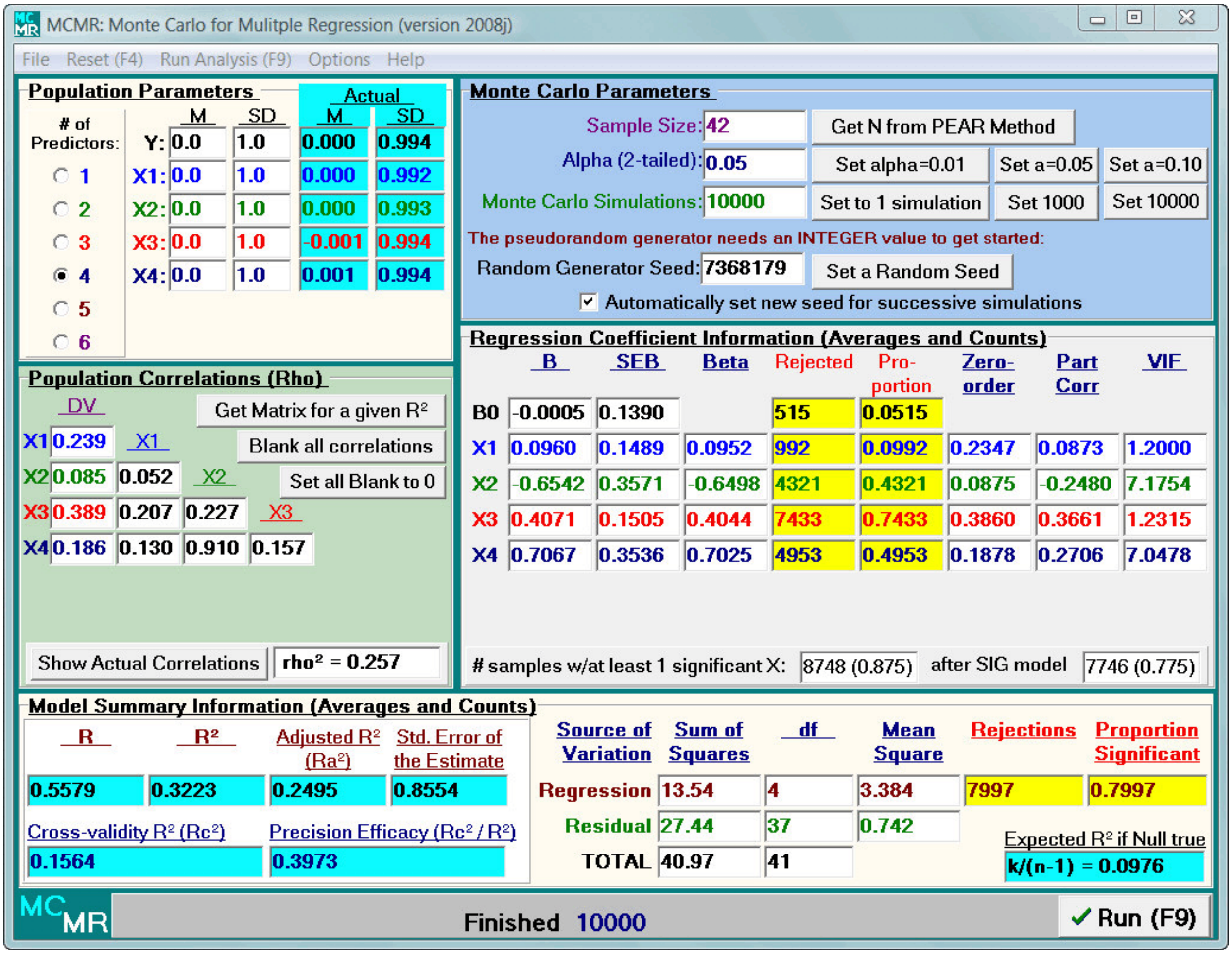
	B	SEB	Beta	Rejected	Proportion	Zero-order	Part Corr	VIF
B0	0.0004	0.1491		483	0.0483			
X1	-0.4518	0.2976	-0.4494	3220	0.3220	0.1687	-0.2192	4.3880
X2	2.2385	0.8839	2.2230	6905	0.6905	0.3242	0.3658	39.2285
X3	-0.9751	0.3966	-0.9700	6668	0.6668	0.1063	-0.3550	7.8630
X4	-0.9337	0.5275	-0.9274	4125	0.4125	0.3119	-0.2562	13.8991

samples w/at least 1 significant X: 7700 (0.770) after SIG model 6544 (0.654)

The most important differences in the results FOR THIS EXAMPLE are the "SEB" and "VIF" results. Note that all SEB values (except for B0) have increased due to the multicollinearity, as have the VIF values.

Other important results, of course, include the regression coefficients ("B" and "Beta") themselves, along with the number of times they were significant. Indeed, different predictors are significant more frequently before (X2 and X4) and after (X2 and X3) due to the multicollinearity introduced into the population, even though the pairwise relationships (zero-order correlations) between the predictors and the dependent variable have not changed.

An Example: Shrinkage and Sample Size



Note that in this example, with a sample size of N = 42 (which provided statistical power for the model of approximately .80), shrinkage occurs from R^2 = .32 down to Adjusted R^2 = .25
or down to Cross-Validity R^2 = .16.

Recall that Adjusted R^2 represents the proportion of variance expected to be accounted for (explained) in the population if this particular regression model is used to predict scores in the population. It is generally considered a better SHRINKAGE estimate when **explanation** is the key purpose for the regression analysis.

Cross-validity R^2 represents the proportion of variance expected to be accounted for if this particular regression model is used in another sample of cases from the same population. It is generally considered a better SHRINKAGE estimate when **prediction** is the key purpose for the regression analysis.

If we use N = 60 (based on 15 cases per predictor), shrinkage is less, but perhaps still too much.	R: 0.5411 R^2: 0.3009 Adjusted R^2 (Ra^2): 0.2501 Std. Error of the Estimate: 0.8578 Cross-validity R^2 (Rc^2): 0.1830 Precision Efficacy (Rc^2 / R^2): 0.5433
If we use, N = 70, which gives us some comfort that Precision Efficacy (using Adjusted R2) will be at least .80, shrinkage is even less.	R: 0.5361 R^2: 0.2946 Adjusted R^2 (Ra^2): 0.2512 Std. Error of the Estimate: 0.8588 Cross-validity R^2 (Rc^2): 0.1937 Precision Efficacy ($AdjR^2/R^2$): 0.8273
If we use N = 150, which gives us comfort that Precision Efficacy (using Cross-validity R2) will be at least .80, reduces shrinkage even further.	R: 0.5205 R^2: 0.2744 Adjusted R^2 (Ra^2): 0.2544 Std. Error of the Estimate: 0.8612 Cross-validity R^2 (Rc^2): 0.2285 Precision Efficacy (Rc^2/R^2): 0.8198

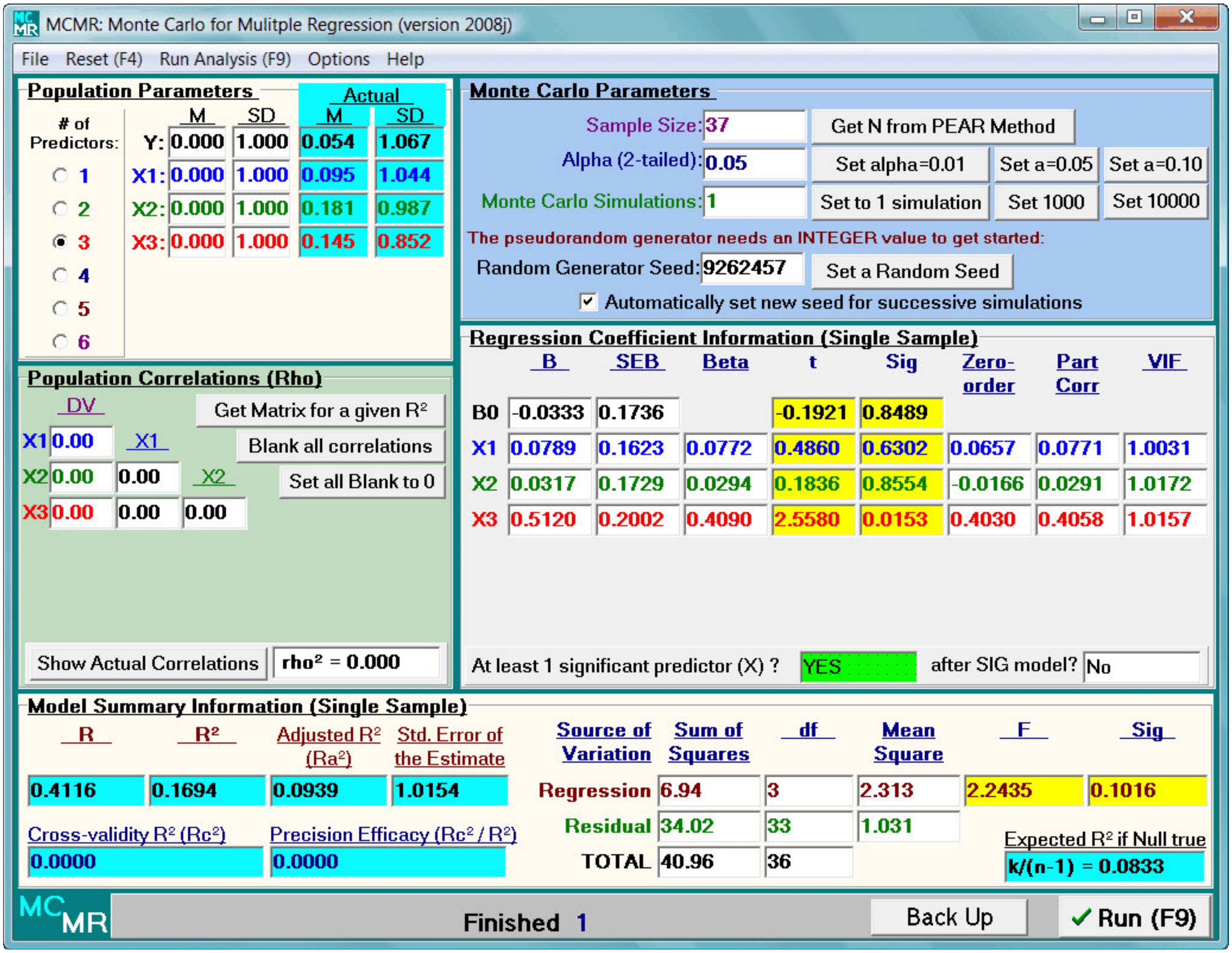
While there is no agreed-upon criterion for SHRINKAGE, several authors have recommended CROSS-VALIDATION as more appropriate methods for determining sample sizes than using statistical power (e.g., Algina & Keselman, 2000; Brooks & Barcikowski, 1999; Park & Dudycha, 1974; Stevens, 1996).

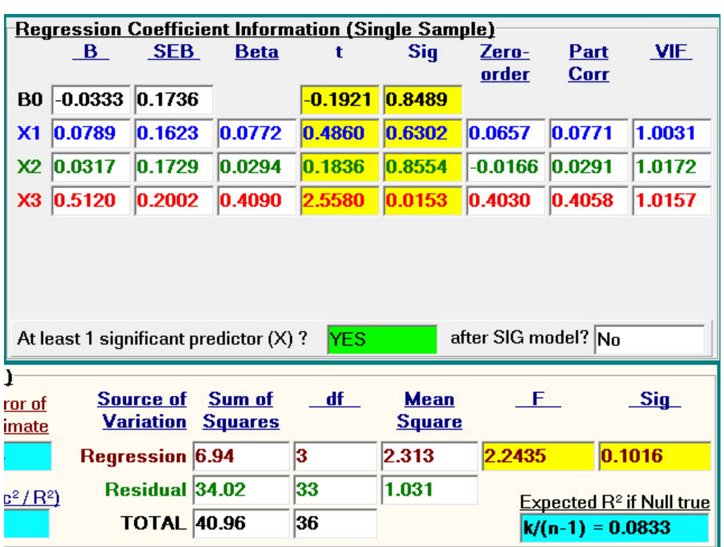
Note that there are also other methods that exist for calculating sample sizes in regression, including statistical power for the *t* tests of the regression coefficients and size of the confidence intervals for the regression coefficients (and therefore size of the standard errors of the regression coefficients).

There are many conventional rules ("rules of thumb") that scholars have recommended over the years as well. These can all be tested and compared using the Monte Carlo method with the MCMR program.

Much more on the topic can be found in Brooks (1998).

An Example: Type I errors (and/or Statistical Power analyses)





We can run SINGLE SAMPLE analyses to show all the possible combinations of Type I errors that occur in multiple regression.

In this first example where all correlations are 0.0, one predictor (X3) is statistically significant, but the model is NOT statistically significant. Therefore, the count boxes show a GREEN YES for “At least 1 significant predictor (X)?” but a white NO for “after SIG model?”

Regression Coefficient Information (Single Sample)

	B	SEB	Beta	t	Sig	Zero-order	Part Corr	VIF
B0	-0.0766	0.2134		-0.3589	0.7219			
X1	0.0122	0.2299	0.0089	0.0530	0.9581	0.0236	0.0088	1.0212
X2	-0.2341	0.1992	-0.2031	-1.1751	0.2484	-0.1393	-0.1953	1.0810
X3	-0.4642	0.2941	-0.2703	-1.5781	0.1241	-0.2209	-0.2623	1.0620

At least 1 significant predictor (X) ? No after SIG model? No

ror of imate	Source of Variation	Sum of Squares	df	Mean Square	F	Sig
	Regression	4.57	3	1.524	1.0643	0.3775
c^2 / R^2)	Residual	47.24	33	1.432		
	TOTAL	51.81	36			

Expected R^2 if Null true
$k/(n-1) = 0.0833$

In this second example where all correlations are 0.0, nothing was statistically significant. This is what we would expect most frequently when the Null Hypothesis is true.

Regression Coefficient Information (Single Sample)

	B	SEB	Beta	t	Sig	Zero-order	Part Corr	VIF
B0	0.1836	0.1607		1.1425	0.2615			
X1	-0.1045	0.1437	-0.1131	-0.7278	0.4719	-0.1698	-0.1110	1.0375
X2	-0.4200	0.1456	-0.4452	-2.8841	0.0069	-0.4331	-0.4400	1.0235
X3	-0.1483	0.1449	-0.1598	-1.0229	0.3138	-0.1242	-0.1561	1.0486

At least 1 significant predictor (X) ? YES after SIG model? YES

ror of imate	Source of Variation	Sum of Squares	df	Mean Square	F	Sig
	Regression	7.77	3	2.591	3.3201	0.0316
c^2 / R^2)	Residual	25.75	33	0.780		
	TOTAL	33.53	36			

Expected R^2 if Null true
$k/(n-1) = 0.0833$

In this third example where all correlations are 0.0, the overall regression model was statistically significant and at least one (here, exactly one, X2) predictor was statistically significant.

Note that different predictors are usually significant in different samples for Robustness (Type I error rate) analyses.

Regression Coefficient Information (Single Sample)

	B	SEB	Beta	t	Sig	Zero-order	Part Corr	VIF
B0	-0.0263	0.1631		-0.1610	0.8730			
X1	0.2982	0.2164	0.2960	1.3777	0.1776	0.3670	0.2092	2.0026
X2	0.6706	0.3353	0.6802	1.9997	0.0538	0.3328	0.3036	5.0203
X3	-0.6156	0.3052	-0.6354	-2.0174	0.0518	0.1506	-0.3063	4.3040

At least 1 significant predictor (X) ? No after SIG model? No

Source of Variation	Sum of Squares	df	Mean Square	F	Sig
Regression	9.22	3	3.074	3.4615	0.0272
Residual	29.30	33	0.888		
TOTAL	38.53	36			

Expected R^2 if Null true $k/(n-1) = 0.0833$

Finished 1 Back Up Run (F9)

NOTE: *This screen comes from an analysis with non-zero correlations, and therefore not a Type I error rate analysis.*

In this fourth example, the overall regression model was statistically significant, but NONE of the predictors was statistically significant. While this appears to be very rare when all correlations are 0.0 (a Type I error rate analysis), it occurs occasionally when the null hypothesis is not true.

Regression Coefficient Information (Averages and Counts)

	B	SEB	Beta	Rejected	Proportion	Zero-order	Part Corr	VIF
B0	-0.0012	0.1705		485	0.0485			
X1	0.0005	0.1747	0.0001	524	0.0524	-0.0002	0.0000	1.0627
X2	-0.0038	0.1739	-0.0032	549	0.0549	-0.0031	-0.0032	1.0617
X3	-0.0006	0.1738	-0.0007	483	0.0483	-0.0008	-0.0008	1.0626

samples w/at least 1 significant X: 1432 (0.143) after SIG model 494 (0.049)

Source of Variation	Sum of Squares	df	Mean Square	Rejections	Proportion Significant
Regression	3.01	3	1.004	518	0.0518
Residual	33.05	33	1.001		
TOTAL	36.06	36			

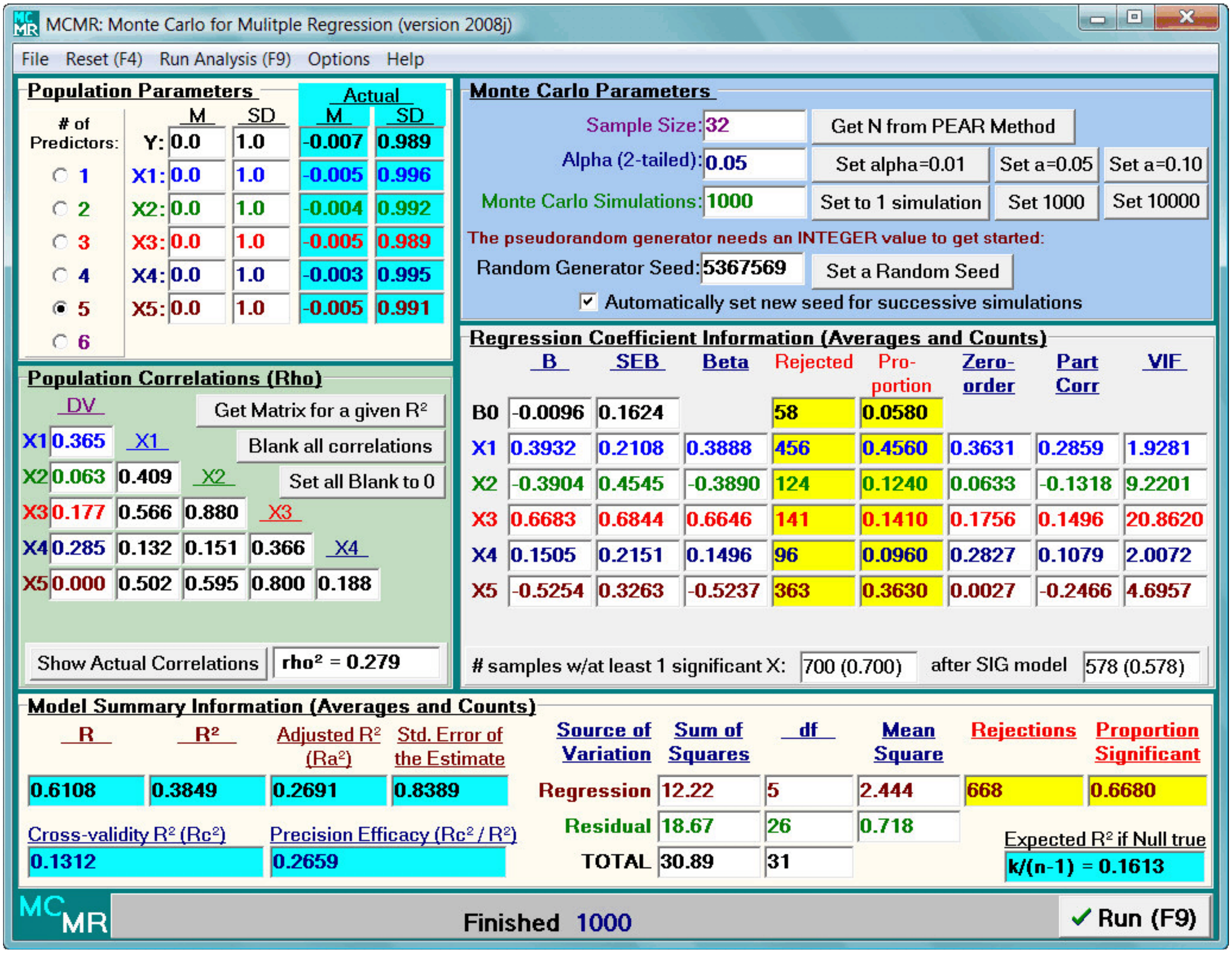
Expected R^2 if Null true $k/(n-1) = 0.0833$

Finally, after running through several samples to show students what a Type I error analysis is like, we can tell them that instead of us going one-by-one through these single samples and keeping track, we can just have the computer do it for us and run 10,000 samples all at once.

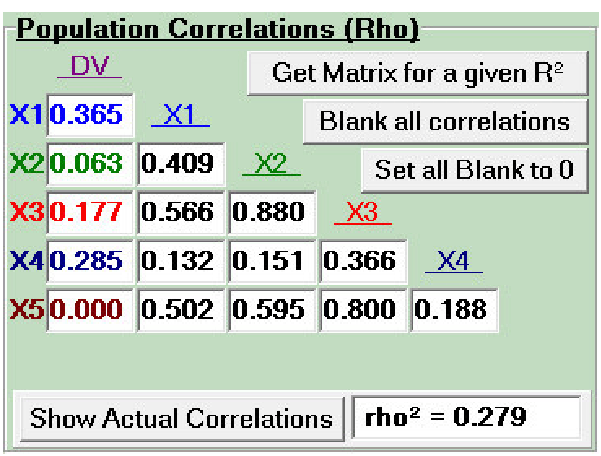
This screen shows the Monte Carlo results for 10,000 simulated samples. One can easily see the approximately .05 Type I error rate expected for all tests.

We can also discuss the idea of a "Protected F" test by reviewing the count boxes. Here, the proportion of simulated samples that had at least one statistically significant predictor FOLLOWING a statistically significant overall regression model is about .049 (5%). However, the proportion of samples that had any number of predictors that were statistically significant was about .14 (14%).

An Example: Suppressor Variables



If we arbitrarily set a population correlation matrix in which one predictor has zero (0.0) correlation with the dependent variable (DV) but has non-zero correlation with the other predictors, we can examine suppressor relationships.



You can see a little better the correlations here.

Note the population multiple rho^2 for this correlation matrix is .279

Model Summary Information (Averages and Counts)

R	R²	Adjusted R² (Ra²)	Std. Error of the Estimate
0.6108	0.3849	0.2691	0.8389

Cross-validity R² (Rc²)	Precision Efficacy (Rc² / R²)
0.1312	0.2659

Source of Variation	Sum of Squares	df	Mean Square	Rejections	Proportion Significant
Regression	12.22	5	2.444	668	0.6680
Residual	18.67	26	0.718		
TOTAL	30.89	31			

Expected R² if Null true: k/(n-1) = 0.1613

We have an R2 value of .38 for this analysis.

Regression Coefficient Information (Averages and Counts)

	B	SEB	Beta	Rejected	Pro-portion	Zero-order	Part Corr	VIF
B0	-0.0096	0.1624		58	0.0580			
X1	0.3932	0.2108	0.3888	456	0.4560	0.3631	0.2859	1.9281
X2	-0.3904	0.4545	-0.3890	124	0.1240	0.0633	-0.1318	9.2201
X3	0.6683	0.6844	0.6646	141	0.1410	0.1756	0.1496	20.8620
X4	0.1505	0.2151	0.1496	96	0.0960	0.2827	0.1079	2.0072
X5	-0.5254	0.3263	-0.5237	363	0.3630	0.0027	-0.2466	4.6957

samples w/at least 1 significant X: 700 (0.700) after SIG model 578 (0.578)

Note the VIF is high for X3, not the variable with 0.0 correlation with the dependent variable (which is X5). However, there is a strong correlation between X3 and X5.

Population Correlations (Rho)

	DV	X1	X2	X3
X1	0.365			
X2	0.063	0.409		
X3	0.177	0.566	0.917	
X4	0.285	0.132	0.151	0.366

Get Matrix for a given R²

Blank all correlations

Set all Blank to 0

Show Actual Correlations

rho² = 0.207

If we remove X5 from the analysis in an effort to remove the multicollinearity (because among the predictors, it has very little correlation with Y), we would have this correlation matrix.

Note that rho^2 is lower without X5 EVEN THOUGH it had no correlation with the Dependent Variable !!

Regression Coefficient Information (Averages and Counts)

	B	SEB	Beta	Rejected	Proportion	Zero-order	Part Corr	VIF
B0	0.0009	0.1680		482	0.0482			
X1	0.4181	0.2190	0.4116	4625	0.4625	0.3588	0.3068	1.8586
X2	-0.0250	0.4111	-0.0244	525	0.0525	0.0637	-0.0099	6.7221
X3	-0.1438	0.4838	-0.1409	605	0.0605	0.1751	-0.0475	9.3599
X4	0.2882	0.2051	0.2838	2760	0.2760	0.2829	0.2258	1.6217

\# samples w/at least 1 significant X: 6035 (0.604) after SIG model 4493 (0.449)

Note that multicollinearity has been removed (as evidenced by all VIF < 10).

Model Summary Information (Averages and Counts)

R	R^2	Adjusted R^2 (Ra^2)	Std. Error of the Estimate
0.5353	0.3011	0.2012	0.8857

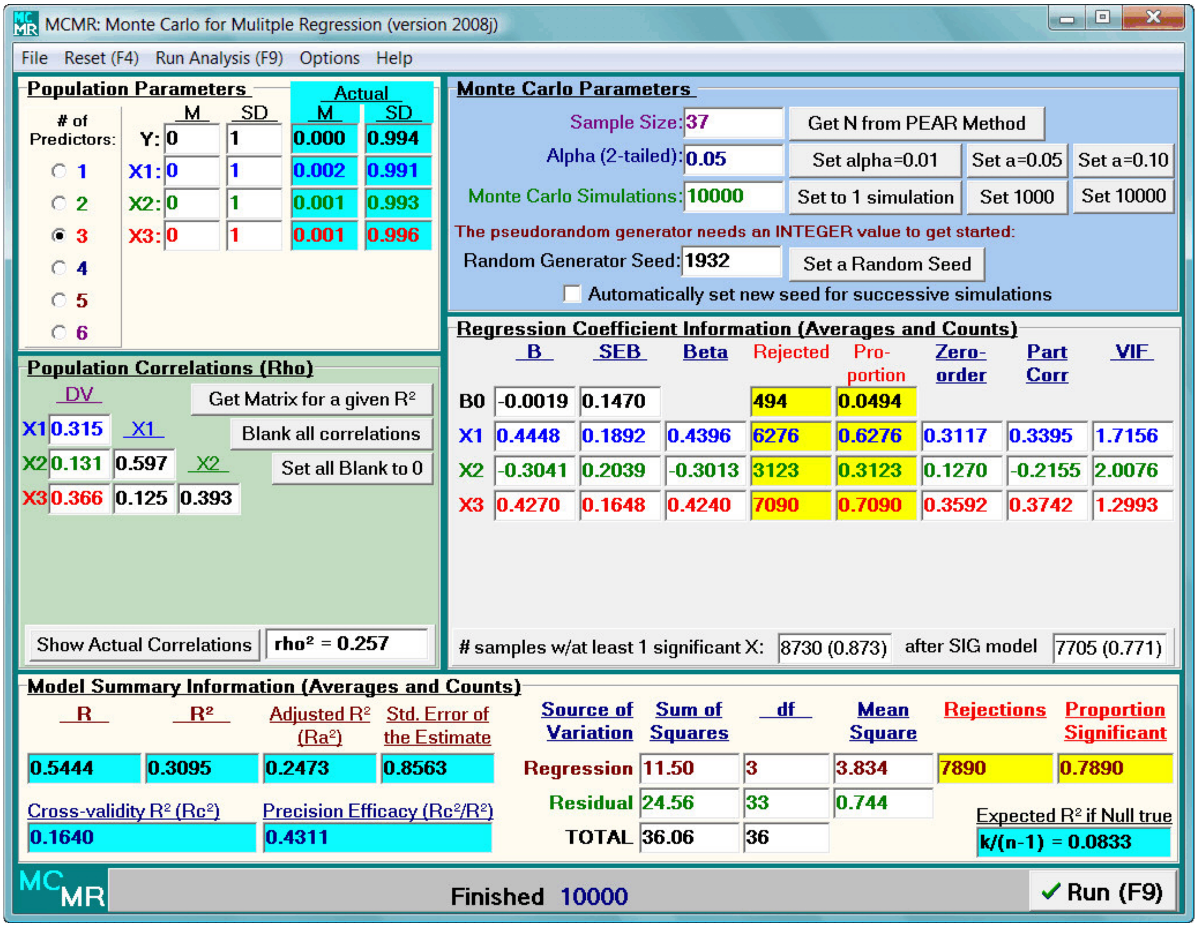
Cross-validity R^2 (Rc^2)	Precision Efficacy (Rc^2 / R^2)
0.0941	0.2184

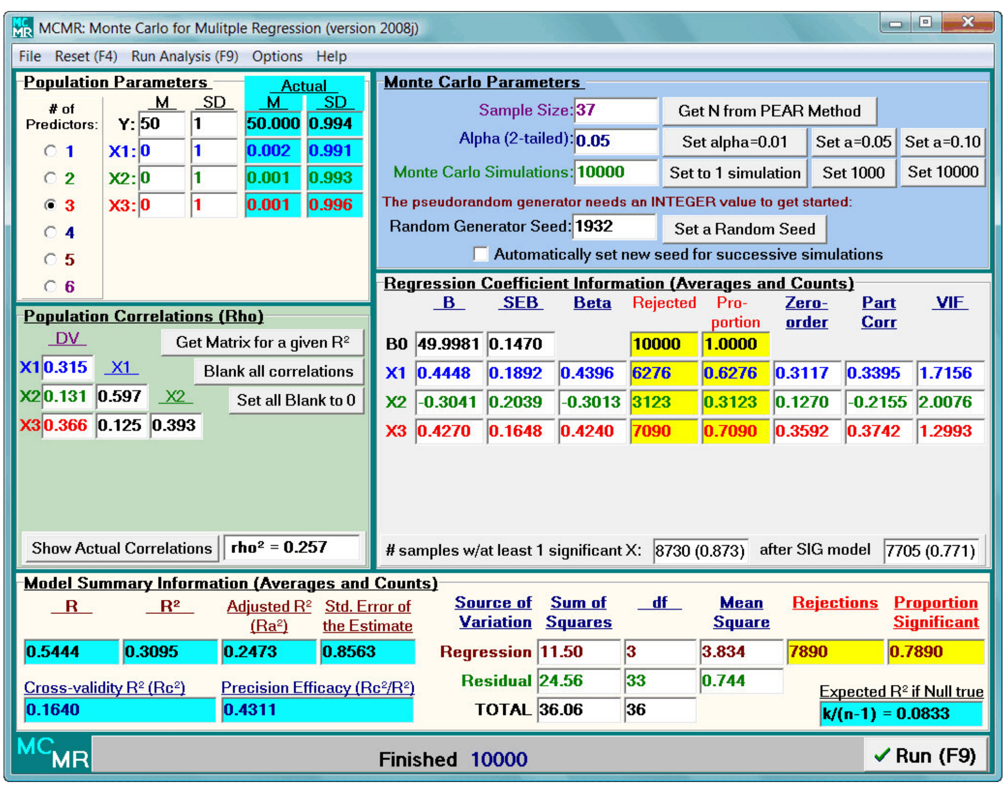
Source of Variation	Sum of Squares	df	Mean Square	Rejections	Proportion Significant
Regression	9.70	4	2.425	5225	0.5225
Residual	21.57	27	0.799		
TOTAL	31.28	31			

Expected R^2 if Null true: $k/(n-1) = 0.1290$

However, the relationship in terms of R^2 is not as high as it was with the apparently useless predictor (.38 then versus .30 now).

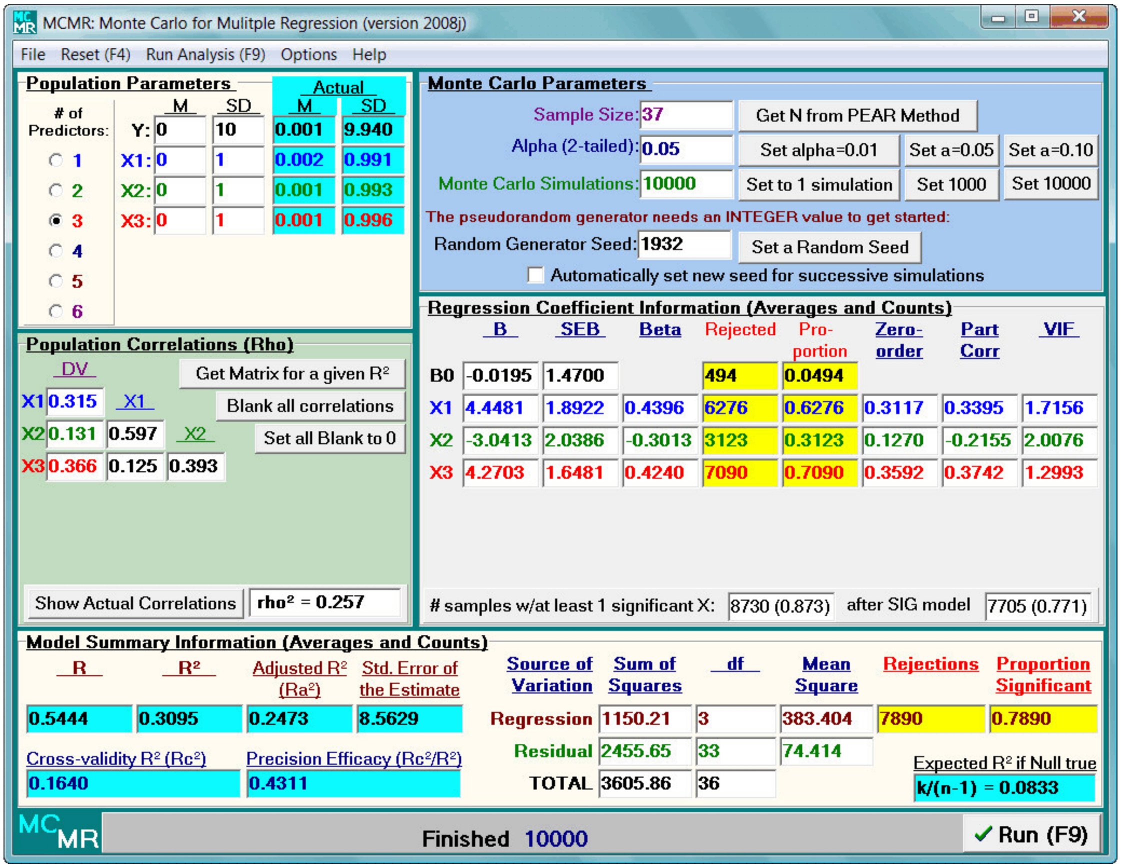
An Example: Impact of Means and Standard Deviations on Regression Results





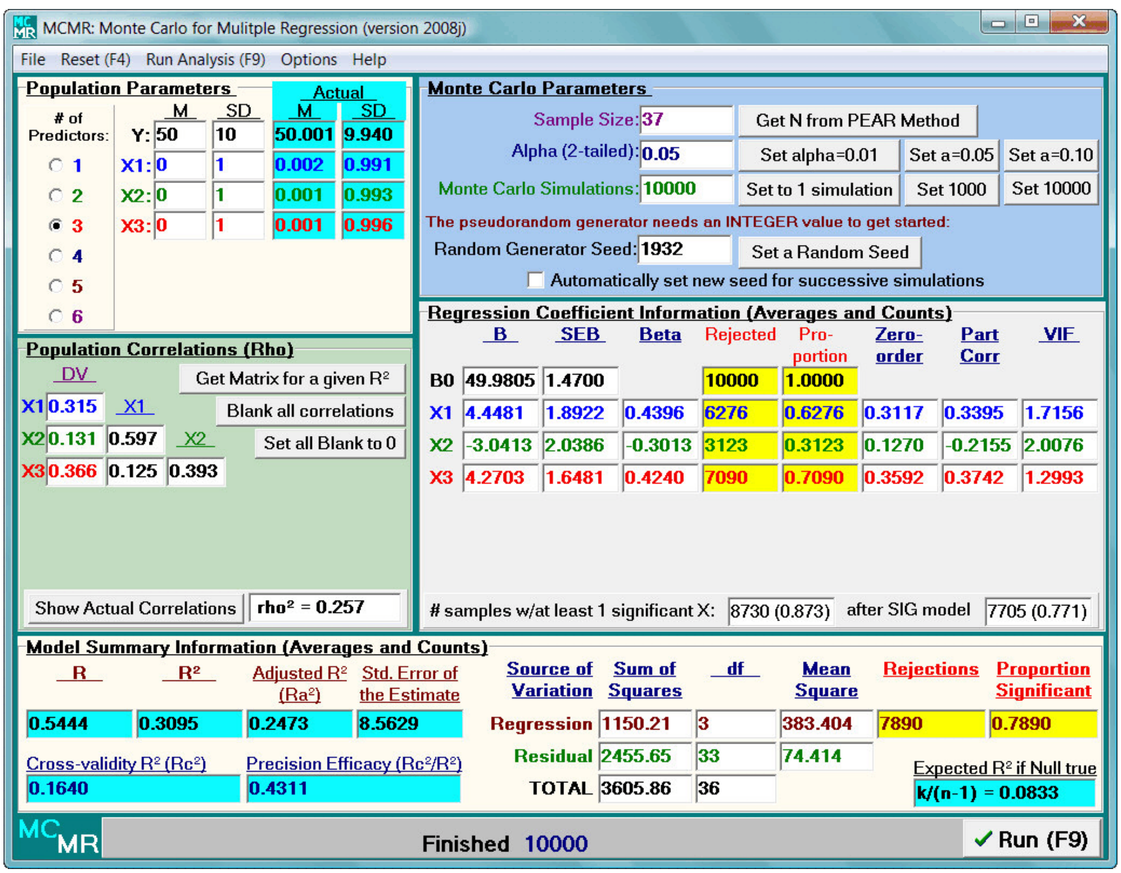
The important thing to notice as we change from all standardized data (above), to a Dependent Variable Mean of 50 (while standard deviation remains 1.0) is that only the CONSTANT B0 and its statistical significance changed.

NOTHING ELSE changed !!

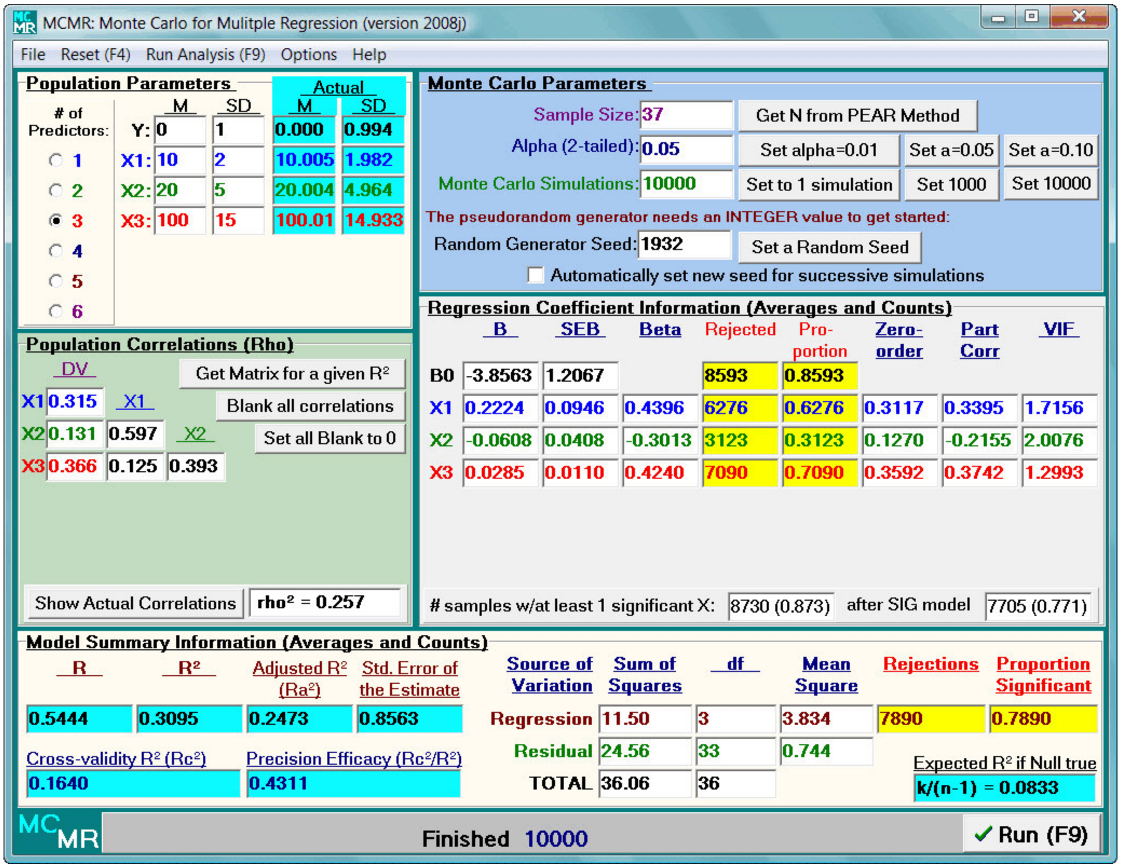


However, when the Dependent Variable Mean is 0.0, but the Standard Deviation changes to 10.0, several things change, most notably the regression coefficients and their significance and the SUMS OF SQUARES.

But none of the other important model information changed (e.g., R^2, F, rejections, Beta, VIF).

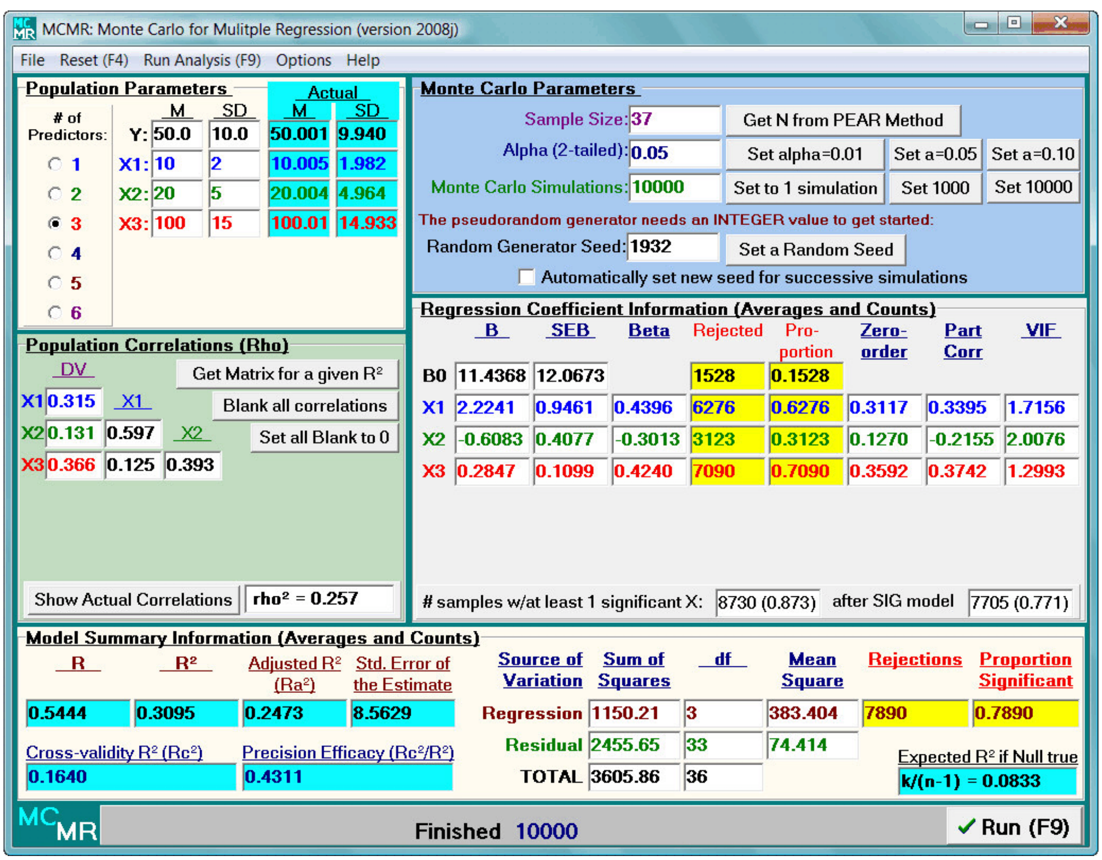


Changing both the Mean and the Standard Deviation combines these previous two results. That is, all the information EXCEPT B0 remains the same as the previous example. But now with the Y mean at 50, B0 changed to match (and is significant more often).



If we change the predictor Means and Standard Deviations, but leave the Dependent Variable Y standardized, you can see several differences — most notably in the regression coefficients.

The "Sum of Squares" values have returned to what they were in the first example.



Finally, if everything changes, the regression coefficients all change, but note that all the MODEL summary information and the CORRELATION information remains the same.

Means and Standard Deviations have not impact on the decisions regarding the Null Hypotheses for either coefficients or the model, nor on the interpretations of the value of the predictors or the model.

References

Algina, J., & Keselman, H. J. (2000). Cross-validation sample sizes. *Applied Psychological Measurement*, *24*, 173–179.

Algina, J., & Olejnik, S. (2000). Determining sample size for accurate estimation of the squared multiple correlation coefficient. *Multivariate Behavioral Research*, *35* (*1*), 119-137.

Brooks, G. P. (1998a, October). *Precision Efficacy Analysis for Regression*. Paper presented at the meeting of the Mid-Western Educational Research Association, Chicago, IL.

Brooks, G. P. (1998b). *Precision efficacy analysis for regression: Development and justification of a new sample size method for multiple linear regression*. Unpublished dissertation, Ohio University, Athens.

Brooks, G. P., & Barcikowski, R. S. (1996). Precision power and its application to the selection of regression sample sizes. *Mid-Western Educational Researcher*, *9*(*4*), 10-17.

Browne, M. W. (1975). Predictive validity of a linear regression equation. *British Journal of Mathematical and Statistical Psychology*, *28*, 79-87.

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd ed.). Hillsdale, NJ: Lawrence Erlbaum Associates.

Darlington, R. B. (1968). Multiple regression in psychological research and practice. *Psychological Bulletin*, *69*, 161-182.

Gatsonis, C., & Sampson, A. R. (1989). Multiple correlation: Exact power and sample size calculations. *Psychological Bulletin*, *106*, 516-524.

Herzberg, P. A. (1969). The parameters of cross-validation. *Psychometrika Monograph Supplement*, *34* (2, Pt. 2).

Kraemer, H. C., & Thiemann, S. (1987). *How many subjects? Statistical power analysis in research*. Newbury Park, CA: Sage.

L'Ecuyer, P. (1988). Efficient and portable combined random number generators. *Communications of the ACM*, *31*, 742-749, 774.

Lord, F. M. (1950). *Efficiency of prediction when a regression equation from one sample is used in a new sample* (Research Bulletin No. 50-40). Princeton, NJ: Educational Testing Service.

Nash, J. C. (1990). *Compact numerical methods for computers: Linear algebra and function minimisation* (2nd ed.). New York: Adam Hilger.

Nicholson, G. E. (1960). Prediction in future samples. In I. Olkin et al. (Eds.), *Contributions to probability and statistics: Essays in honor of Harold Hotelling* (pp. 322-330). Palo Alto, CA: Stanford.

Park, C. N., & Dudycha, A. L. (1974). A cross-validation approach to sample size determination for regression models. *Journal of the American Statistical Association*, *69*, 214-218.

Press, W. H., Flannery, B. P., Teukolsky, S. A., & Vetterling, W. T. (1989). *Numerical recipes in Pascal: The art of scientific computing*. New York: Cambridge University.

Press, W. H., Teukolsky, S. A., Vetterling, W. T., & Flannery, B. P. (1992). *Numerical recipes in FORTRAN: The art of scientific computing* (2nd ed.). New York: Cambridge University.

Rozeboom, W. W. (1978). Estimation of cross-validated multiple correlations: A clarification. *Psychological Bulletin*, *85*, 1348-1351.

Rozeboom, W. W. (1981). The cross-validational accuracy of sample regressions. *Journal of Educational Statistics*, *6*, 179-198.

Sawyer, R. (1982). Sample size and the accuracy of predictions made from multiple regression equations. *Journal of Educational Statistics*, *7*, 91-104.

Stein, C. (1960). Multiple regression. In I. Olkin et al. (Eds.), *Contributions to probability and statistics: Essays in honor of Harold Hotelling* (pp.425-443). Palo Alto, CA: Stanford University.

Stevens, J. (1996). *Applied multivariate statistics for the social sciences* (3rd ed.). Mahwah, NJ: Lawrence Erlbaum Associates.

Uhl, N., & Eisenberg, T. (1970). Predicting shrinkage in the multiple correlation coefficient. *Educational and Psychological Measurement*, *30*, 487-489.

Send correspondence to: Gordon P. Brooks
Ohio University
Email: brooksg@ohio.edu