

A Monte Carlo program for multiple linear regression

Gordon P. Brooks
Ohio University

Abstract

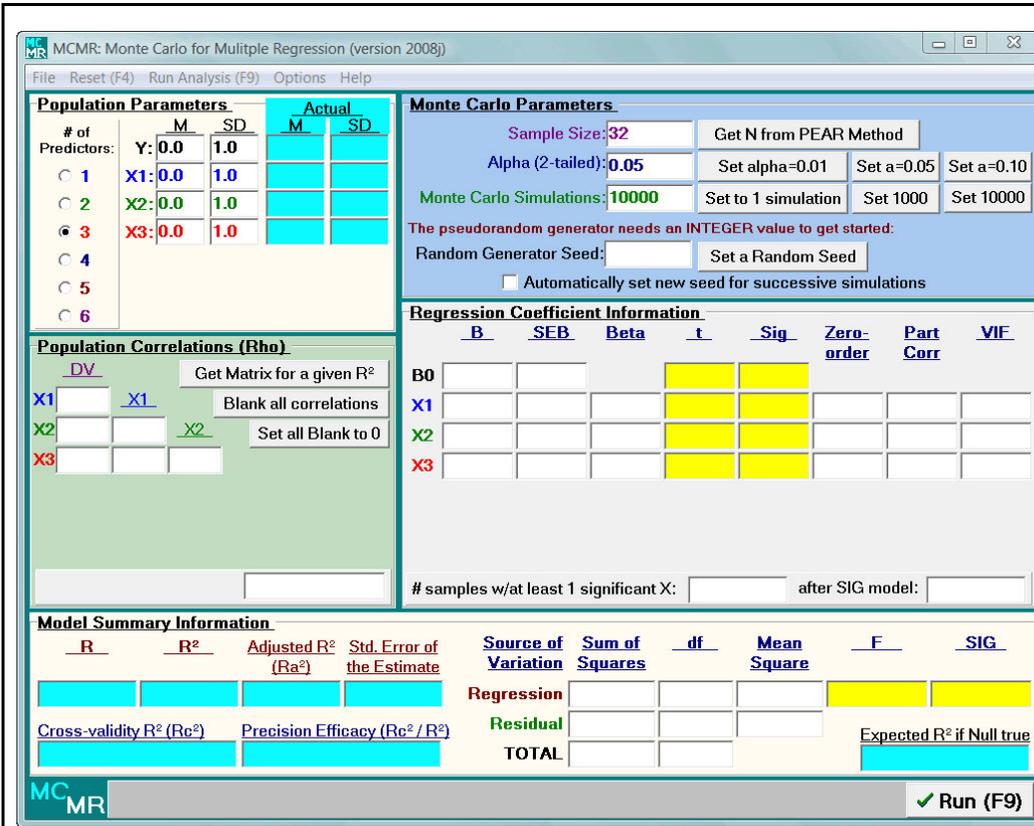
The primary purpose of this presentation is to demonstrate a new computer program that statistics instructors can use to help teach certain regression topics in their courses. In particular, a computer program was written in Borland Delphi 2007 and will run under most recent versions of the Microsoft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

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 through a link at:

<http://oak.cats.ohiou.edu/~brooksg/software.htm>

Paper presented at the annual meeting of the American Educational Research Association, March 24, 2008, New York, NY.

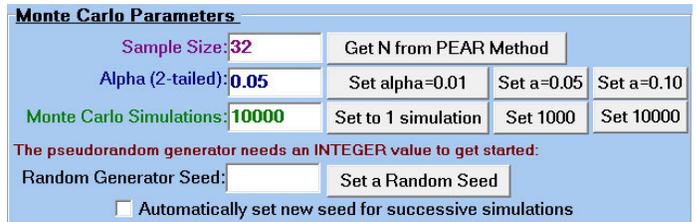


This is the Opening Screen that appears when the program is started (or after the “Reset (F4)” menu option is chosen).

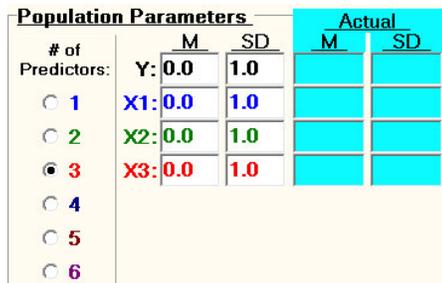
3 sections require user input

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).

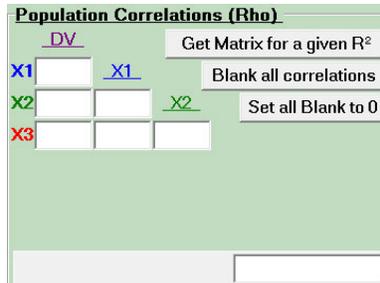
Click “Run” (bottom right) or press F9 to begin the Monte Carlo analysis.



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

MCMR: Monte Carlo for Multiple Regression (version 2008)

File Reset (F4) Run Analysis (F9) Options Help

Population Parameters

# of Predictors:		M		SD	
		Actual		Actual	
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					

Monte Carlo Parameters

Sample Size: 37 Get N from PEAR Method

Alpha (2-tailed): 0.05 Set alpha=0.01 Set a=0.05 Set a=0.10

Monte Carlo Simulations: 10000 Set to 1 simulation Set 1000 Set 10000

The pseudorandom generator needs an INTEGER value to get started:

Random Generator Seed: 1932 Set a Random Seed

Automatically set new seed for successive simulations

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)

Population Correlations (Rho)

DV: Get Matrix for a given R²

X1: 0.0 X1: Blank all correlations

X2: 0.0 X2: 0.0 X2: Set all Blank to 0

X3: 0.0 X3: 0.0 X3:

X4: 0.0 X4: 0.0 X4:

X5: 0.0 X5: 0.0 X5: 0.0

Show Actual Correlations rho² = 0.000

Model Summary Information (Averages and Counts)

R	R ²	Adjusted R ² (Ra ²)	Std. Error of the Estimate	Source of Variation	Sum of Squares	df	Mean Square	Rejections	Proportion Significant
0.3564	0.1385	0.0365	0.9931	Regression	5.01	5	1.001	524	0.0524
				Residual	31.06	31	1.002		
				TOTAL	36.07	36			

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

0.0044 0.0120

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

MC MR Finished 10000 Run (F9)

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		Actual	
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: -0.001 X1:

X2: 0.002 X2: 0.001 X2:

X3: -0.001 X3: -0.001 X3: 0.000 X3:

X4: 0.001 X4: 0.002 X4: 0.000 X4: -0.002 X4:

X5: 0.002 X5: 0.002 X5: -0.001 X5: 0.002 X5: 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)

	<u>B</u>	<u>SEB</u>	<u>Beta</u>	<u>Rejected</u>	<u>Pro- portion</u>	<u>Zero- order</u>	<u>Part Corr</u>	<u>VIF</u>
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)

<u>R</u>	<u>R²</u>	<u>Adjusted R² (Ra²)</u>	<u>Std. Error of the Estimate</u>	<u>Source of Variation</u>	<u>Sum of Squares</u>	<u>df</u>	<u>Mean Square</u>	<u>Rejections</u>	<u>Proportion Significant</u>
0.3564	0.1385	0.0365	0.9931	Regression	5.01	5	1.001	524	0.0524
<u>Cross-validity R² (Rc²)</u>				Residual	31.06	31	1.002	Expected R ² if Null true	
<u>Precision Efficacy (Rc²/R²)</u>				TOTAL	36.07	36	k/(n-1) = 0.1389		
0.0044								0.0120	

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

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

Buttons: 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)

	<u>B</u>	<u>SEB</u>	<u>Beta</u>	<u>t</u>	<u>Sig</u>	<u>Zero-order</u>	<u>Part Corr</u>	<u>VIF</u>
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)

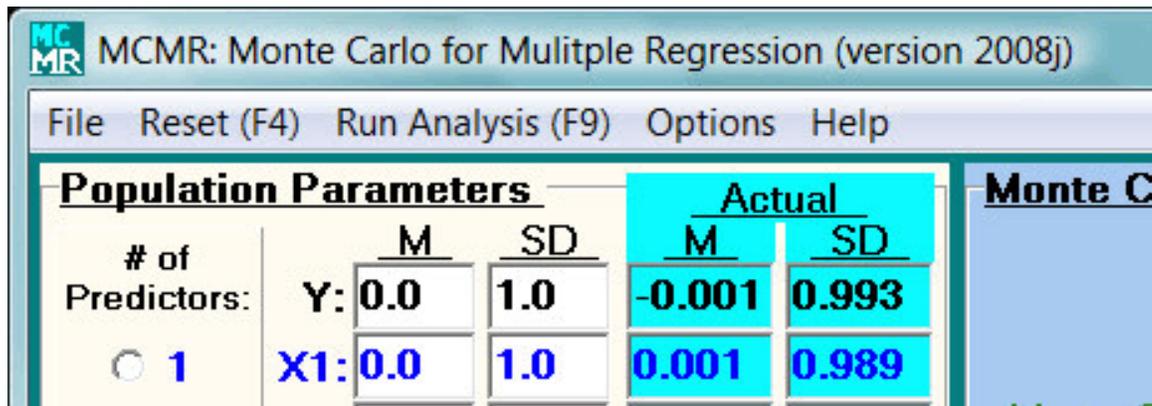
<u>R</u>	<u>R²</u>	<u>Adjusted R² (Ra²)</u>	<u>Std. Error of the Estimate</u>	<u>Source of Variation</u>	<u>Sum of Squares</u>	<u>df</u>	<u>Mean Square</u>	<u>F</u>	<u>Sig</u>
0.4746	0.2253	0.1003	0.9766	Regression	8.60	5	1.719	1.8029	0.1414
<u>Cross-validity R² (Rc²)</u>		<u>Precision Efficacy (Rc²/R²)</u>		Residual	29.57	31	0.954	Expected R ² if Null true	
0.0000		0.0000		TOTAL	38.16	36	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 R² or Cross-validity R² are negative they are set to 0.0 (theoretically, neither they nor R² can be negative).

By the way, the “Expected R² if Null True” box uses the calculation presented by Herzberg (1969), $k/(n-1)$, to show the bias of the R² 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	
View & Save Simulation Data for Models (a little SLOW for maximum 10000 saved)	
View & Save Simulation Data for Predictors (really SLOW for maximum 10000 simulations)	
Import Comma Delimited Data (no missing cases, no case ID, no variable names, DV is first)	
Exit	Ctrl+F4

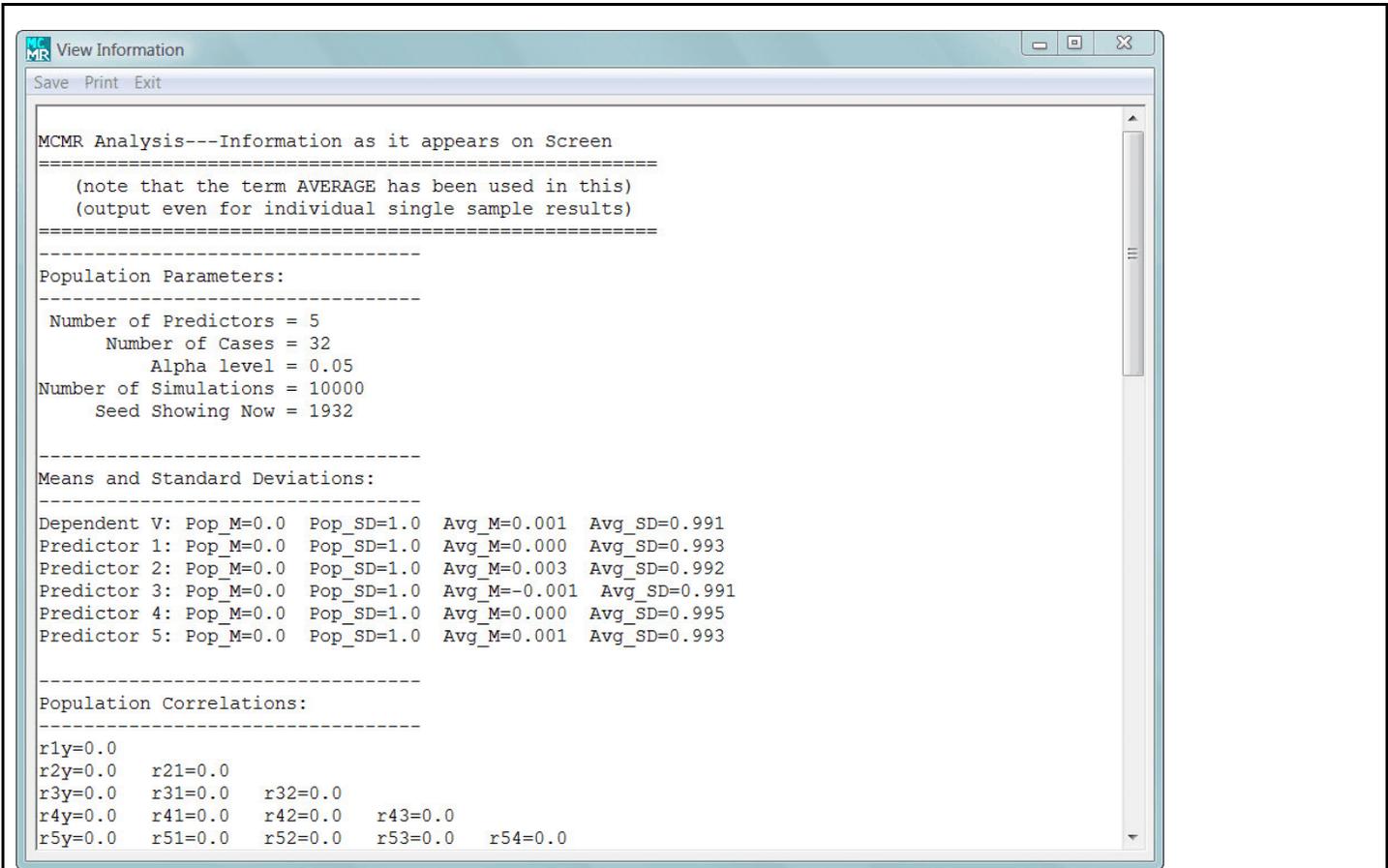
“View and Save Analysis” will show a text version of the results in another window (below), which will also allow you to save and print the results of the analysis.

“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).

ALL OPTIONS START WITH NEXT ANALYSIS

CONSTANT

- Include Constant in Equation
- Do Not Include Constant

EXPECTED R² info (starts with next analysis)

- Always use Expected R² for true Null Hypothesis
- Use Expected R² for given rho² {rho²+ [k/(n-1)](1-rho²)} (Herzberg, 1969)
- Show average shrinkage based on Adjusted R² (R² - Ra²)
- Show average shrinkage based on Cross-Validity R² (R² - Rc²)

PRECISION EFFICACY Calculation

- Use Cross-Validity R² in Precision Efficacy Formula
- Use Adjusted² in Precision Efficacy Formula

CROSS-VALIDITY R² (Shrinkage) FORMULA TO USE:

- Stein (1960)-Darlington (1968) Random-model formula
- Lord (1950) from Uhl & Eisenberg (1970) Random-model formula
- Browne (1975) Random-model formula
- Lord (1950)-Nicholson (1960) Fixed-model formula
- Rozeboom (1978) Fixed-model formula
- Olkin-Pratt from Herzberg (1969) Adjusted R² formula (note: Adjusted R² is from Wherry (1931)-Ezekiel (1930))

"AT LEAST 1" COUNTS

- Include B0 in "At Least 1" counts

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 R² if Null true” — 2 for expected R² and 2 for shrinkage.

Precision Efficacy (Brooks, 1998) is calculated using Cross-Validity R² by default, but could be calculated using Adjusted R². (see help menu for additional information about Precision Efficacy)

Different formulas can be used to calculate Cross-Validity R² — 6 are available here.

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

Precision Efficacy (PEAR) Information

Show Population Regression Equation

User Agreement

About

The “Precision Efficacy (PEAR) Information” option will open a window that contains an excerpt 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.

PRECISION EFFICACY INFORMATION

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

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

where R^2 is the sample coefficient of determination and R_c^2 is the sample cross-validity estimate. For example, if 48% cross-validity shrinkage from sample $R^2 = .50$ to $R_c^2 = .26$ 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

Print Done

The screenshot shows a pop-up window titled 'mcmr2008' with the regression equation: $Zy\text{-hat} = -0.8572 \cdot Zx1 + 0.5020 \cdot Zx2 + -0.5561 \cdot Zx3 + 0.5027 \cdot Zx4 + 1.0254 \cdot Zx5$. Below the window, the 'Regression Coefficient Information' table is visible:

	B	SEB	Beta
B0	-0.0014	0.1519	
X1	-0.8649	0.4010	-0.8590
X2	0.5065	0.2512	0.5034
X3	-0.5611	0.3209	-0.5575
X4	0.5031	0.2228	0.4999
X5	1.0332	0.4063	1.0249

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 contains the following text:

USER AGREEMENT

Carefully read the following User Agreement (License, Terms of Use, and Disclaimer of Warranty).

Use of the MCMR.EXE (Monte Carlo for Multiple Regression) software program provided with this Agreement constitutes acceptance of these terms and conditions of use. If you do not agree to the terms of this agreement, do not use the MCMR software program.

LICENSE

MCMR is a copyrighted program and is NOT public domain. The user is granted license, not ownership, to use the MCMR software program on any computer subject to the restrictions described in the User Agreement and Disclaimer.

MCMR is Freeware. The user is licensed to make an unlimited number of exact copies of the MCMR software program, to give these exact copies to any other person for their personal use, and to distribute the MCMR software program in its unmodified form only via disk, email, or local area network.

If these methods of distribution are unavailable, any person wanting to use the MCMR software program should be directed either to contact the author or to visit the author's internet web site (the URL is provided below and may be posted on any web site).

Close

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

The 'About...' window displays the following information:

MC MR MCMR

MCMR: Monte Carlo for Multiple Regression (version 2008j)

Brooks, G. P. (2008, March). A Monte Carlo program for multiple linear regression. Paper presented at the 2008 meeting of the American Educational Research Association, New York, NY.

Copyright © 2008 Gordon P. Brooks
 Contact: brooksg@ohiou.edu

Web Site URL: <http://oak.cats.ohiou.edu/~brooksg/software.htm>

Close

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

Population Correlations (Rho)

DV Get Matrix for a given R²

X1 **0.170** X1 Blank all correlations

X2 **0.329** **0.725** X2 Set all Blank to 0

X3 **0.107** **0.631** **0.810** X3

X4 **0.318** **0.413** **0.849** **0.476**

Show Actual Correlations **rho² = 0.260**

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.

Set R²

What R² would you like for your data?

(please note that because this will serve as a POPULATION matrix, the sample data may not produce this matrix exactly)

How close do you want to approximate this R² value?

(please note that the closer you wish to approximate R², the longer this process may take --- 0.01 works relatively well)

What VIF value do you consider problematic?

(many scholars consider it problematic when VIF is over 10, but some consider it troublesome even when VIF > 5)

Approximately how many **NEGATIVE** correlations do you want in your population matrix?

NONE Some About HALF Most ALL

(please note that this is based on probability, so you may not get exactly the right number---you can always try again)

How much **MULTICOLLINEARITY** would you like built into your population correlation matrix?

Absolutely None (all correlations among predictor = 0)

No worrisome Collinearity (no VIF values for any predictor above the "problematic" value set above)

1 or 2 predictor with VIF over the "problematic" value set above

2 or 3 predictors with VIF over the "problematic" value set above

4 or more predictors with VIF over the "problematic" value set above

Set one predictor correlation with DV to be 0.0

(please note that some options may not work well with some numbers of predictors)

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.

What R² would you like for your data?

(please note that because this will serve as a POPULATION matrix, the sample data may not produce this matrix exactly)

You can choose any R² for your POPULATION correlation matrix (so really this is a rho² or ρ² 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 R² values calculated in the samples (other than they should be from the population with the derived population correlation matrix).

How close do you want to approximate this R² value?

(please note that the closer you wish to approximate R², the longer this process may take --- 0.01 works relatively well)

You can choose how close you want to approximate the population R² 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 R² in the POPULATION correlation matrix, and says nothing about the samples drawn during the Monte Carlo process.

What VIF value do you consider problematic?

(many scholars consider it problematic when VIF is over 10, but some consider it troublesome even when VIF > 5)

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 $Tolerance = 1 - R_j^2$, where R_j^2 is the squared correlation when the j^{th} predictor acts as a temporary dependent variable being predicted by all the other predictors.

Approximately how many NEGATIVE correlations do you want in your population matrix?

 NONE

 Some

 About HALF

 Most

 ALL

(please note that this is based on probability, so you may not get exactly the right number---you can always try again)

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

How much MULTICOLLINEARITY would you like built into your population correlation matrix?

- Absolutely None (all correlations among predictor = 0)
- No worrisome Collinearity (no VIF values for any predictor above the "problematic" value set above)
- 1 or 2 predictor with VIF over the "problematic" value set above
- 2 or 3 predictors with VIF over the "problematic" value set above
- 4 or more predictors with VIF over the "problematic" value set above
- Set one predictor correlation with DV to be 0.0

(please note that some options may not work well with some numbers of predictors)

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.

Set one predictor correlation with DV to be 0.0
please note that some options may not work well with some

Stop



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).

mcmr2008

This matrix is not Positive Definite, as correlation matrices are assumed to be. We need a good correlation matrix in order to generate data. Please try another matrix.

OK

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

$$\epsilon = R - R_C^2$$

whereas Proportional Shrinkage (PS) might be written as

$$PS = (R^2 - R_C^2) / R^2$$

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

$$PE = R_C^2 / R^2$$

Solving $PE = 1 - \epsilon/R^2$ for ϵ , and replacing R^2 with an expected, a priori R_E^2 , results in the formula

$$\epsilon = R_E^2 - (PE)(R_E^2)$$

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 $\epsilon = .05$ or $\epsilon = .10$.

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

$$N = \frac{p(1 - \rho^2) + \varepsilon}{\varepsilon},$$

where

$$\varepsilon = (R^2 - R_A^2),$$

as compared to

$$N = \frac{(k+1)(2 - 2\rho^2 + \varepsilon)}{\varepsilon},$$

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

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

$$\varepsilon = (1 - PE)\rho^2$$

where it would simplify to

$$\varepsilon = .2\rho^2$$

for $PE = .80$ (just like it would also for the Cross-Validity approach).

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 ρ^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

MCMR: Monte Carlo for Multiple Regression (version 2008)

File Reset (F4) Run Analysis (F9) Options Help

Population Parameters

# of Predictors:	Y:	M		SD		Actual	
		M	SD	M	SD	M	SD
1	X1:	0.0	1.0	0.000	0.993	0.000	0.994
2	X2:	0.0	1.0	0.001	0.993	0.001	0.993
3	X3:	0.0	1.0	0.002	0.995	0.002	0.995
4	X4:	0.0	1.0	0.000	0.993	0.000	0.993

Monte Carlo Parameters

Sample Size: 37 Get N from PEAR Method

Alpha (2-tailed): 0.05 Set alpha=0.01 Set a=0.05 Set a=0.10

Monte Carlo Simulations: 10000 Set to 1 simulation Set 1000 Set 10000

The pseudorandom generator needs an INTEGER value to get started:

Random Generator Seed: 1932 Set a Random Seed

Automatically set new seed for successive simulations

Regression Coefficient Information (Averages and Counts)

	B	SEB	Beta	Rejected	Pro-portion	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)

Population Correlations (Rho)

DV: X1, X2, X3, X4

Get Matrix for a given R²

Blank all correlations

Set all Blank to 0

Show Actual Correlations rho² = 0.250

Model Summary Information (Averages and Counts)

R	R ²	Adjusted R ² (Ra ²)	Std. Error of the Estimate	Source of Variation	Sum of Squares	df	Mean Square	Rejections	Proportion Significant
0.5576	0.3233	0.2399	0.8595	Regression	11.95	4	2.988	7131	0.7131
				Residual	24.02	32	0.750		
				TOTAL	35.97	36			

Cross-validity R² (Rc²): 0.1375 Precision Efficacy (Rc² / R²): 0.3320

Expected R² if Null true k/(n-1) = 0.1111

MC MR Finished 10000 Run (F9)

Let's assume that all predictor intercorrelations are 0.0, while predictor correlations with Y are non-zero such that rho² = .25.

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

Population Correlations (Rho)

DV: X1, X2, X3, X4

Get Matrix for a given R²

Blank all correlations

Set all Blank to 0

Show Actual Correlations rho² = 0.250

Regression Coefficient Information (Averages and Counts)

	<u>B</u>	<u>SEB</u>	<u>Beta</u>	<u>Rejected</u>	<u>Pro- portion</u>	<u>Zero- order</u>	<u>Part Corr</u>	<u>VIF</u>
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 rho² 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 R² 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 rho².

Population Correlations (Rho)

Get Matrix for a given R²

Blank all correlations

Set all Blank to 0

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

Show Actual Correlations rho² = 0.260

Regression Coefficient Information (Averages and Counts)

	<u>B</u>	<u>SEB</u>	<u>Beta</u>	<u>Rejected</u>	<u>Pro- portion</u>	<u>Zero- order</u>	<u>Part Corr</u>	<u>VIF</u>
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

MCMR: Monte Carlo for Multiple Regression (version 2008)

File Reset (F4) Run Analysis (F9) Options Help

Population Parameters

# of Predictors:	Y:	M		SD		Actual	
		M	SD	M	SD	M	SD
1	X1:	0.0	1.0	0.000	0.994	0.000	0.992
2	X2:	0.0	1.0	0.000	0.993	0.000	0.993
3	X3:	0.0	1.0	-0.001	0.994	-0.001	0.994
4	X4:	0.0	1.0	0.001	0.994	0.001	0.994

Monte Carlo Parameters

Sample Size: 42 Get N from PEAR Method

Alpha (2-tailed): 0.05 Set alpha=0.01 Set a=0.05 Set a=0.10

Monte Carlo Simulations: 10000 Set to 1 simulation Set 1000 Set 10000

The pseudorandom generator needs an INTEGER value to get started:

Random Generator Seed: 7368179 Set a Random Seed

Automatically set new seed for successive simulations

Regression Coefficient Information (Averages and Counts)

	B	SEB	Beta	Rejected	Pro-portion	Zero-order	Part Corr	VIF
B0	-0.0005	0.1390		515	0.0515			
X1	0.0960	0.1489	0.0952	992	0.0992	0.2347	0.0873	1.2000
X2	-0.6542	0.3571	-0.6498	4321	0.4321	0.0875	-0.2480	7.1754
X3	0.4071	0.1505	0.4044	7433	0.7433	0.3860	0.3661	1.2315
X4	0.7067	0.3536	0.7025	4953	0.4953	0.1878	0.2706	7.0478

samples w/at least 1 significant X: 8748 (0.875) after SIG model 7746 (0.775)

Population Correlations (Rho)

DV: Get Matrix for a given R²

X1: 0.239 X1: Blank all correlations

X2: 0.085 0.052 X2: Set all Blank to 0

X3: 0.389 0.207 0.227 X3:

X4: 0.186 0.130 0.910 0.157

Show Actual Correlations rho² = 0.257

Model Summary Information (Averages and Counts)

R	R ²	Adjusted R ² (Ra ²)	Std. Error of the Estimate	Source of Variation	Sum of Squares	df	Mean Square	Rejections	Proportion Significant
0.5579	0.3223	0.2495	0.8554	Regression	13.54	4	3.384	7997	0.7997
				Residual	27.44	37	0.742		
				TOTAL	40.97	41			

Cross-validity R² (Rc²): 0.1564 Precision Efficacy (Rc² / R²): 0.3973

Expected R² if Null true k/(n-1) = 0.0976

MC MR Finished 10000 Run (F9)

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.

<p>If we use N = 60 (based on 15 cases per predictor), shrinkage is less, but perhaps still too much.</p>	<table border="1"> <thead> <tr> <th><u>R</u></th> <th><u>R²</u></th> <th><u>Adjusted R² (Ra²)</u></th> <th><u>Std. Error of the Estimate</u></th> </tr> </thead> <tbody> <tr> <td>0.5411</td> <td>0.3009</td> <td>0.2501</td> <td>0.8578</td> </tr> <tr> <td colspan="2"><u>Cross-validity R² (Rc²)</u></td> <td colspan="2"><u>Precision Efficacy (Rc² / R²)</u></td> </tr> <tr> <td colspan="2">0.1830</td> <td colspan="2">0.5433</td> </tr> </tbody> </table>	<u>R</u>	<u>R²</u>	<u>Adjusted R² (Ra²)</u>	<u>Std. Error of the Estimate</u>	0.5411	0.3009	0.2501	0.8578	<u>Cross-validity R² (Rc²)</u>		<u>Precision Efficacy (Rc² / R²)</u>		0.1830		0.5433	
<u>R</u>	<u>R²</u>	<u>Adjusted R² (Ra²)</u>	<u>Std. Error of the Estimate</u>														
0.5411	0.3009	0.2501	0.8578														
<u>Cross-validity R² (Rc²)</u>		<u>Precision Efficacy (Rc² / R²)</u>															
0.1830		0.5433															
<p>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.</p>	<table border="1"> <thead> <tr> <th><u>R</u></th> <th><u>R²</u></th> <th><u>Adjusted R² (Ra²)</u></th> <th><u>Std. Error of the Estimate</u></th> </tr> </thead> <tbody> <tr> <td>0.5361</td> <td>0.2946</td> <td>0.2512</td> <td>0.8588</td> </tr> <tr> <td colspan="2"><u>Cross-validity R² (Rc²)</u></td> <td colspan="2"><u>Precision Efficacy (AdjR²/R²)</u></td> </tr> <tr> <td colspan="2">0.1937</td> <td colspan="2">0.8273</td> </tr> </tbody> </table>	<u>R</u>	<u>R²</u>	<u>Adjusted R² (Ra²)</u>	<u>Std. Error of the Estimate</u>	0.5361	0.2946	0.2512	0.8588	<u>Cross-validity R² (Rc²)</u>		<u>Precision Efficacy (AdjR²/R²)</u>		0.1937		0.8273	
<u>R</u>	<u>R²</u>	<u>Adjusted R² (Ra²)</u>	<u>Std. Error of the Estimate</u>														
0.5361	0.2946	0.2512	0.8588														
<u>Cross-validity R² (Rc²)</u>		<u>Precision Efficacy (AdjR²/R²)</u>															
0.1937		0.8273															
<p>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.</p>	<table border="1"> <thead> <tr> <th><u>R</u></th> <th><u>R²</u></th> <th><u>Adjusted R² (Ra²)</u></th> <th><u>Std. Error of the Estimate</u></th> </tr> </thead> <tbody> <tr> <td>0.5205</td> <td>0.2744</td> <td>0.2544</td> <td>0.8612</td> </tr> <tr> <td colspan="2"><u>Cross-validity R² (Rc²)</u></td> <td colspan="2"><u>Precision Efficacy (Rc²/R²)</u></td> </tr> <tr> <td colspan="2">0.2285</td> <td colspan="2">0.8198</td> </tr> </tbody> </table>	<u>R</u>	<u>R²</u>	<u>Adjusted R² (Ra²)</u>	<u>Std. Error of the Estimate</u>	0.5205	0.2744	0.2544	0.8612	<u>Cross-validity R² (Rc²)</u>		<u>Precision Efficacy (Rc²/R²)</u>		0.2285		0.8198	
<u>R</u>	<u>R²</u>	<u>Adjusted R² (Ra²)</u>	<u>Std. Error of the Estimate</u>														
0.5205	0.2744	0.2544	0.8612														
<u>Cross-validity R² (Rc²)</u>		<u>Precision Efficacy (Rc²/R²)</u>															
0.2285		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)

Population Parameters

# of Predictors:	Y:	M		SD		Actual	
		M	SD	M	SD		
1	X1:	0.000	1.000	0.054	1.067	0.095	1.044
2	X2:	0.000	1.000	0.181	0.987		
3	X3:	0.000	1.000	0.145	0.852		

Monte Carlo Parameters

Sample Size: 37 (Get N from PEAR Method)

Alpha (2-tailed): 0.05 (Set alpha=0.01, Set a=0.05, Set a=0.10)

Monte Carlo Simulations: 1 (Set to 1 simulation, Set 1000, Set 10000)

The pseudorandom generator needs an INTEGER value to get started:
 Random Generator Seed: 9262457 (Set a Random Seed)

Automatically set new seed for successive simulations

Regression Coefficient Information (Single Sample)

	B	SEB	Beta	t	Sig	Zero-order	Part Corr	VIF
B0	-0.0333	0.1736		-0.1921	0.8489			
X1	0.0789	0.1623	0.0772	0.4860	0.6302	0.0657	0.0771	1.0031
X2	0.0317	0.1729	0.0294	0.1836	0.8554	-0.0166	0.0291	1.0172
X3	0.5120	0.2002	0.4090	2.5580	0.0153	0.4030	0.4058	1.0157

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

Model Summary Information (Single Sample)

R	R ²	Adjusted R ² (Ra ²)	Std. Error of the Estimate	Source of Variation	Sum of Squares	df	Mean Square	F	Sig
0.4116	0.1694	0.0939	1.0154	Regression	6.94	3	2.313	2.2435	0.1016
				Residual	34.02	33	1.031		
				TOTAL	40.96	36			

Cross-validity R² (Rc²): 0.0000
 Precision Efficacy (Rc² / R²): 0.0000
 Expected R² if Null true k/(n-1) = 0.0833

MC MR Finished 1 Back Up Run (F9)

Regression Coefficient Information (Single Sample)

	B	SEB	Beta	t	Sig	Zero-order	Part Corr	VIF
B0	-0.0333	0.1736		-0.1921	0.8489			
X1	0.0789	0.1623	0.0772	0.4860	0.6302	0.0657	0.0771	1.0031
X2	0.0317	0.1729	0.0294	0.1836	0.8554	-0.0166	0.0291	1.0172
X3	0.5120	0.2002	0.4090	2.5580	0.0153	0.4030	0.4058	1.0157

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

Model Summary Information (Single Sample)

Source of Variation	Sum of Squares	df	Mean Square	F	Sig
Regression	6.94	3	2.313	2.2435	0.1016
Residual	34.02	33	1.031		
TOTAL	40.96	36			

Expected R² if Null true k/(n-1) = 0.0833

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) ? after SIG model?

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.

Source of Variation	Sum of Squares	df	Mean Square	F	Sig
Regression	4.57	3	1.524	1.0643	0.3775
Residual	47.24	33	1.432		
TOTAL	51.81	36			

Expected R² if Null true
k/(n-1) = 0.0833

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) ? after SIG model?

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.

Source of Variation	Sum of Squares	df	Mean Square	F	Sig
Regression	7.77	3	2.591	3.3201	0.0316
Residual	25.75	33	0.780		
TOTAL	33.53	36			

Expected R² if Null true
k/(n-1) = 0.0833

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 R² 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	Pro-portion	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 R² 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

Population Parameters

# of Predictors:	Y:	M		SD		Actual	
		M	SD	M	SD	M	SD
1	X1:	0.0	1.0	-0.007	0.989	-0.005	0.996
2	X2:	0.0	1.0	-0.004	0.992		
3	X3:	0.0	1.0	-0.005	0.989		
4	X4:	0.0	1.0	-0.003	0.995		
5	X5:	0.0	1.0	-0.005	0.991		

Monte Carlo Parameters

Sample Size: 32
 Alpha (2-tailed): 0.05
 Monte Carlo Simulations: 1000
 Random Generator Seed: 5367569

Population Correlations (Rho)

DV	X1	X2	X3	X4	X5
0.365					
0.063	0.409				
0.177	0.566	0.880			
0.285	0.132	0.151	0.366		
0.000	0.502	0.595	0.800	0.188	

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

Model Summary Information (Averages and Counts)

R	R ²	Adjusted R ² (Ra ²)	Std. Error of the Estimate	Source of Variation	Sum of Squares	df	Mean Square	Rejections	Proportion Significant
0.6108	0.3849	0.2691	0.8389	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

Finished 1000 Run (F9)

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.

Population Correlations (Rho)

DV	X1	X2	X3	X4	X5
0.365					
0.063	0.409				
0.177	0.566	0.880			
0.285	0.132	0.151	0.366		
0.000	0.502	0.595	0.800	0.188	

rho² = 0.279

You can see a little better the correlations here.

Note the population multiple rho² for this correlation matrix is .279

Model Summary Information (Averages and Counts)									
<u>R</u>	<u>R²</u>	<u>Adjusted R²</u> (Ra ²)	<u>Std. Error of</u> <u>the Estimate</u>	<u>Source of</u> <u>Variation</u>	<u>Sum of</u> <u>Squares</u>	<u>df</u>	<u>Mean</u> <u>Square</u>	<u>Rejections</u>	<u>Proportion</u> <u>Significant</u>
0.6108	0.3849	0.2691	0.8389	Regression	12.22	5	2.444	668	0.6680
Cross-validity R ² (Rc ²)		Precision Efficacy (Rc ² / R ²)		Residual	18.67	26	0.718	Expected R ² if Null true k/(n-1) = 0.1613	
0.1312		0.2659		TOTAL	30.89	31			

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

Regression Coefficient Information (Averages and Counts)								
	<u>B</u>	<u>SEB</u>	<u>Beta</u>	<u>Rejected</u>	<u>Pro-</u> <u>portion</u>	<u>Zero-</u> <u>order</u>	<u>Part</u> <u>Corr</u>	<u>VIF</u>
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)				
<u>DV</u>	Get Matrix for a given R ²			
X1	0.365	X1	Blank all correlations	
X2	0.063	0.409	X2	Set all Blank to 0
X3	0.177	0.566	0.917	X3
X4	0.285	0.132	0.151	0.366

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 rho² is lower without X5 EVEN THOUGH it had no correlation with the Dependent Variable !!

Regression Coefficient Information (Averages and Counts)

	<u>B</u>	<u>SEB</u>	<u>Beta</u>	<u>Rejected</u>	<u>Pro- portion</u>	<u>Zero- order</u>	<u>Part Corr</u>	<u>VIF</u>
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)

<u>R</u>	<u>R²</u>	<u>Adjusted R² (Ra²)</u>	<u>Std. Error of the Estimate</u>	<u>Source of Variation</u>	<u>Sum of Squares</u>	<u>df</u>	<u>Mean Square</u>	<u>Rejections</u>	<u>Proportion Significant</u>
0.5353	0.3011	0.2012	0.8857	Regression	9.70	4	2.425	5225	0.5225
<u>Cross-validity R² (Rc²)</u>		<u>Precision Efficacy (Rc² / R²)</u>		Residual	21.57	27	0.799	Expected R ² if Null true	
0.0941		0.2184		TOTAL	31.28	31		k/(n-1) = 0.1290	

However, the relationship in terms of R² 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

MCMR: Monte Carlo for Multiple Regression (version 2008j)

File Reset (F4) Run Analysis (F9) Options Help

Population Parameters		Actual	
# of Predictors	M	SD	
Y:	0	1	0.000 0.994
X1:	0	1	0.002 0.991
X2:	0	1	0.001 0.993
X3:	0	1	0.001 0.996

Population Correlations (Rho)	
DV	
X1	0.315
X2	0.131
X3	0.366

Monte Carlo Parameters	
Sample Size:	37
Alpha (2-tailed):	0.05
Monte Carlo Simulations:	10000
Random Generator Seed:	1932

Regression Coefficient Information (Averages and Counts)							
	B	SEB	Beta	Rejected	Proportion	Zero-order	VIF
B0	-0.0019	0.1470		494	0.0494		
X1	0.4448	0.1892	0.4396	6276	0.6276	0.3117	0.3395
X2	-0.3041	0.2039	-0.3013	3123	0.3123	0.1270	-0.2155
X3	0.4270	0.1648	0.4240	7090	0.7090	0.3592	0.3742

Model Summary Information (Averages and Counts)							
R	R ²	Adjusted R ² (Ra ²)	Std. Error of the Estimate	Source of Variation	Sum of Squares	df	Mean Square
0.5444	0.3095	0.2473	0.8563	Regression	11.50	3	3.834
				Residual	24.56	33	0.744
				TOTAL	36.06	36	

Rejections: 7890, Proportion Significant: 0.7890

Expected R² if Null true: k/(n-1) = 0.0833

Finished 10000

MCMR: Monte Carlo for Multiple Regression (version 2008j)

File Reset (F4) Run Analysis (F9) Options Help

Population Parameters		Actual	
# of Predictors	M	SD	
Y:	50	1	50.000 0.994
X1:	0	1	0.002 0.991
X2:	0	1	0.001 0.993
X3:	0	1	0.001 0.996

Population Correlations (Rho)	
DV	
X1	0.315
X2	0.131
X3	0.366

Monte Carlo Parameters	
Sample Size:	37
Alpha (2-tailed):	0.05
Monte Carlo Simulations:	10000
Random Generator Seed:	1932

Regression Coefficient Information (Averages and Counts)							
	B	SEB	Beta	Rejected	Proportion	Zero-order	VIF
B0	49.9981	0.1470		10000	1.0000		
X1	0.4448	0.1892	0.4396	6276	0.6276	0.3117	0.3395
X2	-0.3041	0.2039	-0.3013	3123	0.3123	0.1270	-0.2155
X3	0.4270	0.1648	0.4240	7090	0.7090	0.3592	0.3742

Model Summary Information (Averages and Counts)							
R	R ²	Adjusted R ² (Ra ²)	Std. Error of the Estimate	Source of Variation	Sum of Squares	df	Mean Square
0.5444	0.3095	0.2473	0.8563	Regression	11.50	3	3.834
				Residual	24.56	33	0.744
				TOTAL	36.06	36	

Rejections: 7890, Proportion Significant: 0.7890

Expected R² if Null true: k/(n-1) = 0.0833

Finished 10000

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 !!

MCMR: Monte Carlo for Multiple Regression (version 2008)

File Reset (F4) Run Analysis (F9) Options Help

Population Parameters

# of Predictors	Y:	M		SD	
		M	SD	M	SD
1	X1: 0	0	1	0.001	0.991
2	X2: 0	0	1	0.001	0.993
3	X3: 0	0	1	0.001	0.996

Monte Carlo Parameters

Sample Size: 37 Get N from PEAR Method

Alpha (2-tailed): 0.05 Set alpha=0.01 Set a=0.05 Set a=0.10

Monte Carlo Simulations: 10000 Set to 1 simulation Set 1000 Set 10000

The pseudorandom generator needs an INTEGER value to get started:

Random Generator Seed: 1932 Set a Random Seed

Automatically set new seed for successive simulations

Regression Coefficient Information (Averages and Counts)

	B	SEB	Beta	Rejected	Pro-portion	Zero-order	Part Corr	VIF
B0	-0.0195	1.4700		494	0.0494			
X1	4.4481	1.8922	0.4396	6276	0.6276	0.3117	0.3395	1.7156
X2	-3.0413	2.0386	-0.3013	3123	0.3123	0.1270	-0.2155	2.0076
X3	4.2703	1.6481	0.4240	7090	0.7090	0.3592	0.3742	1.2993

samples w/at least 1 significant X: 8730 (0.873) after SIG model 7705 (0.771)

Model Summary Information (Averages and Counts)

R	R ²	Adjusted R ² (Ra ²)	Std. Error of the Estimate	Source of Variation	Sum of Squares	df	Mean Square	Rejections	Proportion Significant
0.5444	0.3095	0.2473	8.5629	Regression	1150.21	3	383.404	7890	0.7890
				Residual	2455.65	33	74.414		
				TOTAL	3605.86	36			

Cross-validity R² (Rc²): 0.1640 Precision Efficacy (Rc²/R²): 0.4311

Expected R² if Null true k/(n-1) = 0.0833

Finished 10000 Run (F9)

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., R², F, rejections, Beta, VIF).

MCMR: Monte Carlo for Multiple Regression (version 2008)

File Reset (F4) Run Analysis (F9) Options Help

Population Parameters

# of Predictors	Y:	M		SD	
		M	SD	M	SD
1	X1: 0	0	1	0.002	0.991
2	X2: 0	0	1	0.001	0.993
3	X3: 0	0	1	0.001	0.996

Monte Carlo Parameters

Sample Size: 37 Get N from PEAR Method

Alpha (2-tailed): 0.05 Set alpha=0.01 Set a=0.05 Set a=0.10

Monte Carlo Simulations: 10000 Set to 1 simulation Set 1000 Set 10000

The pseudorandom generator needs an INTEGER value to get started:

Random Generator Seed: 1932 Set a Random Seed

Automatically set new seed for successive simulations

Regression Coefficient Information (Averages and Counts)

	B	SEB	Beta	Rejected	Pro-portion	Zero-order	Part Corr	VIF
B0	49.9805	1.4700		10000	1.0000			
X1	4.4481	1.8922	0.4396	6276	0.6276	0.3117	0.3395	1.7156
X2	-3.0413	2.0386	-0.3013	3123	0.3123	0.1270	-0.2155	2.0076
X3	4.2703	1.6481	0.4240	7090	0.7090	0.3592	0.3742	1.2993

samples w/at least 1 significant X: 8730 (0.873) after SIG model 7705 (0.771)

Model Summary Information (Averages and Counts)

R	R ²	Adjusted R ² (Ra ²)	Std. Error of the Estimate	Source of Variation	Sum of Squares	df	Mean Square	Rejections	Proportion Significant
0.5444	0.3095	0.2473	8.5629	Regression	1150.21	3	383.404	7890	0.7890
				Residual	2455.65	33	74.414		
				TOTAL	3605.86	36			

Cross-validity R² (Rc²): 0.1640 Precision Efficacy (Rc²/R²): 0.4311

Expected R² if Null true k/(n-1) = 0.0833

Finished 10000 Run (F9)

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).

Population Parameters

# of Predictors	Y:	M		SD		Actual	
		M	SD	M	SD		
1	X1:	10	2	10.005	1.982		
2	X2:	20	5	20.004	4.964		
3	X3:	100	15	100.01	14.933		

Monte Carlo Parameters

Sample Size: 37
 Alpha (2-tailed): 0.05
 Monte Carlo Simulations: 10000
 Random Generator Seed: 1932

Regression Coefficient Information (Averages and Counts)

	B	SEB	Beta	Rejected	Proportion	Zero-order	Part Corr	VIF
B0	-3.8563	1.2067		8593	0.8593			
X1	0.2224	0.0946	0.4396	6276	0.6276	0.3117	0.3395	1.7156
X2	-0.6068	0.4048	-0.3013	3123	0.3123	0.1270	-0.2155	2.0076
X3	0.0285	0.0110	0.4240	7090	0.7090	0.3592	0.3742	1.2993

Model Summary Information (Averages and Counts)

R	R ²	Adjusted R ² (Ra ²)	Std. Error of the Estimate	Source of Variation	Sum of Squares	df	Mean Square	Rejections	Proportion Significant
0.5444	0.3095	0.2473	0.8563	Regression	11.50	3	3.834	7890	0.7890
				Residual	24.56	33	0.744		
				TOTAL	36.06	36			

Cross-validity R² (Rc²): 0.1640
 Precision Efficacy (Rc²/R²): 0.4311
 Expected R² if Null true: k/(n-1) = 0.0833

Finished 10000

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.

Population Parameters

# of Predictors	Y:	M		SD		Actual	
		M	SD	M	SD		
1	X1:	10	2	10.005	1.982		
2	X2:	20	5	20.004	4.964		
3	X3:	100	15	100.01	14.933		

Monte Carlo Parameters

Sample Size: 37
 Alpha (2-tailed): 0.05
 Monte Carlo Simulations: 10000
 Random Generator Seed: 1932

Regression Coefficient Information (Averages and Counts)

	B	SEB	Beta	Rejected	Proportion	Zero-order	Part Corr	VIF
B0	11.4368	12.0673		1528	0.1528			
X1	2.2241	0.9461	0.4396	6276	0.6276	0.3117	0.3395	1.7156
X2	-0.6083	0.4077	-0.3013	3123	0.3123	0.1270	-0.2155	2.0076
X3	0.2847	0.1099	0.4240	7090	0.7090	0.3592	0.3742	1.2993

Model Summary Information (Averages and Counts)

R	R ²	Adjusted R ² (Ra ²)	Std. Error of the Estimate	Source of Variation	Sum of Squares	df	Mean Square	Rejections	Proportion Significant
0.5444	0.3095	0.2473	8.5629	Regression	1150.21	3	383.404	7890	0.7890
				Residual	2455.65	33	74.414		
				TOTAL	3605.86	36			

Cross-validity R² (Rc²): 0.1640
 Precision Efficacy (Rc²/R²): 0.4311
 Expected R² if Null true: k/(n-1) = 0.0833

Finished 10000

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 impacted 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

(including references cited within the MCMR program itself)

- 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 doctoral 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 University.
- 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-validated 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.