Short Communication

A fitter use of Monte Carlo simulations in regression models

Alessandro Ferrarini
Department of Evolutionary and Functional Biology, University of Parma, Via G. Saragat 4, I-43100 Parma, Italy
E-mail: sgtpm@libero.it, alessandro.ferrarini@unipr.it

Received 17 August 2011; *Accepted 22 September 2011*; *Published online 1 December 2011*
IAEES

Abstract
In this article, I focus on the use of Monte Carlo simulations (MCS) within regression models, being this application very frequent in biology, ecology and economy as well. I'm interested in enhancing a typical fault in this application of MCS, i.e. the inner correlations among independent variables are not used when generating random numbers that fit their distributions. By means of an illustrative example, I provide proof that the misuse of MCS in regression models produces misleading results. Furthermore, I also provide a solution for this topic.

Keywords Monte Carlo simulations; regression models; correlated independent variables.

1 Introduction

The Monte Carlo method was invented in the 1940s by John von Neumann, Stanislaw Ulam and Nicholas Metropolis, when they were working at the Los Alamos National Laboratory (Metropolis and Ulam, 1949; Eckhardt, 1987). Thenceforward, the Monte Carlo method has been frequently employed in many scientific fields (Fishman, 1995; Hammersley and Handscomb, 1975; Rubinstein and Kroese, 2007).

In this article, I focus on the use of Monte Carlo simulations (MCS) in regression modelling, which is a very frequent application in biology, ecology, medicine and economy as well. I'm interested in enhancing a typical fault in this use of MCS, i.e. the correlations among independent variables are not used when generating random numbers that fit their distributions. The misuse of MCS in regression models can lead to misleading results. In this view, I provide a solution for this topic.

2 The MCS in Regression Modelling

The typical use of MCS within regression models in ecology, biology and economy is as follows:

a) a deterministic relation between Y and ($X_1...X_n$) is found;

b) the distributions of $X_1...X_n$ are fitted from sampled data along with distribution parameters;

c) for each independent variable $X_1...X_n$, *m* random numbers are simulated that follow the distribution properties of the previous step. As a result, a (*m* rows * *n* columns) matrix of simulated independent data is available;

d) the deterministic relation between Y and ($X_1...X_n$) is applied to the previous matrix. As a result, *m* simulated values for Y are generated;

e) the resulting distribution for Y is examined for reasonable assumptions about the nature of uncertainty expected from the model and source data.

3 How It Should Be

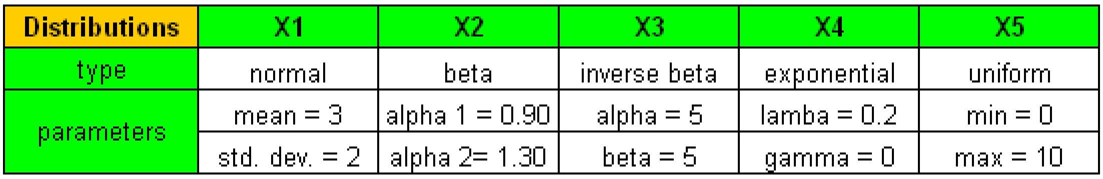
The previous framework neglects to use a pivotal information present within the source data, i.e. the correlations existing among independent variables. This is a inner property of sampled data that can't be ignored, because it's at least as meaningful in representing the system under study as the other parameters considered in the MCS of regression models.

If the simulated (*m* rows * *n* columns) matrix of simulated independent data does not possess the same correlations as the starting sampled data, hence it does not represent the system under study and the resulting outcomes about the real nature of Y are not correct. Hence two further steps are required within the previous framework: correlations among $X_1...X_n$ must be calculated, and then integrated within the simulation of *m* random numbers for $X_1...X_n$.

4 An Illustrative Example

Let's suppose that $X_1...X_n$ have distributions as depicted in Table 1.

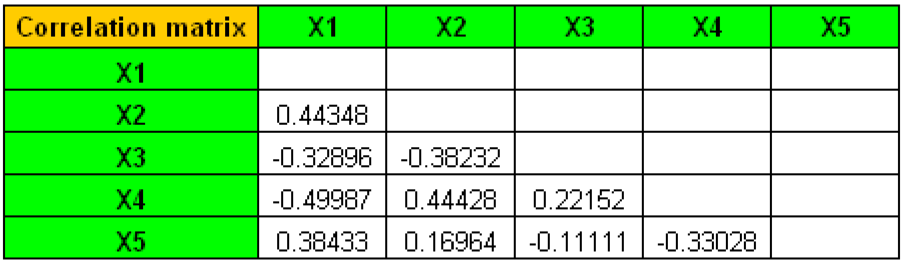
Table 1 Supposed distributions of independent variables X1...X5



Distributions	X1	X2	X3	X4	X5
type	normal	beta	inverse beta	exponential	uniform
parameters	mean = 3	alpha 1 = 0.90	alpha = 5	lamba = 0.2	min = 0
	std. dev. = 2	alpha 2= 1.30	beta = 5	gamma = 0	max = 10

In addition, let's suppose that correlations exist among variables as showed in Table 2. Since 4 variables out of 5 are not Gaussian-like distributed, a Spearman's non-parametric coefficient for continuous data is the most correct.

Table 2 Supposed correlations among independent variables X1...X5



Correlation matrix	X1	X2	X3	X4	X5
X1					
X2	0.44348				
X3	-0.32896	-0.38232			
X4	-0.49987	0.44428	0.22152		
X5	0.38433	0.16964	-0.11111	-0.33028	

Last, let's say to have found a regression model as follows:

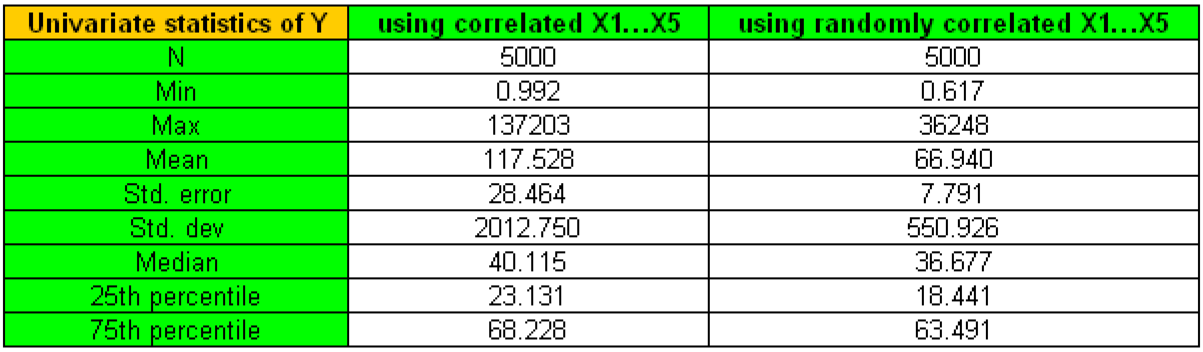
$$Y = 2.3*X_1^2 + 6.4*X_2*X_3 + 7.2*X_4/X_5 \qquad (1)$$

Now, I'll follow two ways. Firstly, I'll generate 5000 values for $X_1 ... X_n$ just following distribution properties of Table 1 (i.e., typical use of MCS in regression models). Secondly, I'll generate 5000 values for $X_1 ... X_n$ that also take into account correlations of Table 2 among variables. See online supplementary material for the Excel spreadsheets of both approaches.

To do this, I've used the algorithm by Iman and Conover (1982) to generate sets of intercorrelated random deviates. The paper of Iman and Conover is highly recommended reading. To this aim, I've coined an Excel worksheet named CorrMCS that applies all the previously-depicted MCM steps along with the algorithm by Iman and Conover. CorrMCS is able to perform both the typical use of MCS in regression modelling, and the improved approach proposed here to better fit the system under study. When applied to the previous example, results are listed in Table 3.

It's clear that using correlated or uncorrelated MCS produced different results, both with regard to Y's predicted range and mean-modal values.

Table 3 Results for Y basing on a) correlated, and b) randomly correlated Monte Carlo simulations



Univariate statistics of Y	using correlated X1...X5	using randomly correlated X1...X5
N	5000	5000
Min	0.992	0.617
Max	137203	36248
Mean	117.528	66.940
Std. error	28.464	7.791
Std. dev	2012.750	550.926
Median	40.115	36.677
25th percentile	23.131	18.441
75th percentile	68.228	63.491

I add several hints:

a) the improved MCS approach is as necessary as the number of independent variables increases;

b) I suggest to use only significant correlations ($p<0.05$ or $p<0.01$) among $X_1 ... X_n$ when using the improved MCS approach; nonsignificant correlations should be set to 0 when generating Monte Carlo numbers;

c) when generating Monte Carlo numbers, attention should be paid to the kind of correlations among $X_1 ... X_n$. Pearson's *r* is correct only for Gaussian-like distributed sampled data.

5 Conclusions

The misuse of MCS in regression models leads to very serious errors in ecology, biology, medicine and economy. I've developed the computer framework CorrMCS based on Iman and Conover's algorithm to overcome these tough mistakes. I provide consultancy to any research or working groups that need to correctly apply MCS in their experiments.

References

Eckhardt R. 1987. Stan Ulam, John von Neumann, and the Monte Carlo method. Los Alamos Science, 15: 131–137

Fishman GS. 1995. Monte Carlo: Concepts, Algorithms, and Applications. Springer, New York, USA

Hammersley JM, Handscomb D.C. 1975. Monte Carlo Methods. Methuen, London, UK

Metropolis N, Ulam S. 1949. The Monte Carlo Method. Journal of the American Statistical Association, 44: 335–341

Iman RL, Conover W.J. 1982. A distribution-free approach to inducing rank correlation among input variables. Communication in Statistics - Simulation and Computation, 11: 311-334

Rubinstein RY, Kroese D.P. 2007. Simulation and the Monte Carlo Method (2nd edition). John Wiley & Sons, New York, USA