Statistical Simulations Using the CALIS Procedure

David P. MacKinnon

The purpose of this paper is to describe how to conduct statistical simulations of covariance structure models using SAS BASE, SAS STAT and the new CALIS program. Other covariance structure programs like LISREL (Joreskog & Sorbom, 1989) and EQS (Bentler, 1989) have statistical simulation capabilities. In this respect, CALIS is not as useful as these other programs. However, as shown below the SAS program can be used to conduct general statistical simulations for covariance structure models. The program consists of generating data based on the regression equations for the hypothesized model in the OATA step, estimating the model in CALIS and then repeating this step for a number of replications to determine the sampling variability of simulation statistics. The method is illustrated for a manifest variable mediation model, a one factor measurement model, and a three factor mediation model with multiple indicators.

Introduction

Covariance structure analysis is used to develop and test comprehensive models of the relationships among variables. The method includes measurement models for hypothesized, unobserved or latent constructs and a structural model for the relationships among the constructs. Covariance structure analysis is also known as causal modeling, linear structural relations, and structural equation modeling. It includes factor analysis, multiple regression and path analysis as special cases.

Covariance structure analysis is now widely used in social and behavioral sciences, including psychology (Bentler, 1983; McArdle & Epstein, 1987) and sociology (Sobel, 1982; 1986; Winship & Mare, 1983). Several covariance structure analysis computer programs are available such as EQS (Bentler, 1989), LISREL (Joreskog & Sorbom, 1988), and LINCS (Schoenberg, 1987). Recently, SAS introduced a covariance structure analysis program called CALIS that combines many aspects of these original programs. For the data analyst, the CALIS program is especially powerful because SAS data sets are easily accessed for covariance structure analysis.

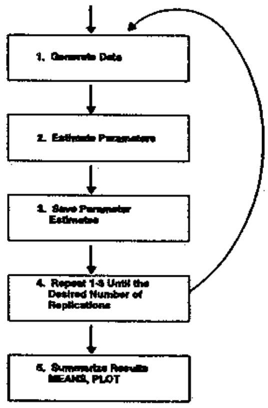
Statistical simulation studies have been an integral part of the development and application of covariance structure modeling. Statistical simulations have been used to examine violations of model assumptions and to develop new statistics to assess the adequacy of the models (Boomsma, 1985; Muthen & Kaplan, 1985). Many of the computer programs described above include statistical simulation capabilities. CALIS does not have a statistical simulation routine in the program. However, as shown below, it is relatively straightforward to conduct such simulations in SAS using the DATA step to generate data, the CALIS program to estimate the models and other SAS capabilities such as MACRO to set up the replications.

The purpose of this paper is to introduce statistical simulations with the CALIS program. The substantive area for two of the simulations is the estimation of mediated effects. Statistical simulations are used to evaluate the bias in several statistics.

CALSIM Program

The CALis SIMulation program consists of five major steps: (1) generate data, (2) estimate parameters, (3) save parameter estimates and standard errors, (4) repeat the generation and estimation for the desired number of replications and (5) summarize the results of the simulation. The loop of replications within the simulations is accomplished using the %DO, %END, and %MEND statements in the MACRO procedure. At the end of the program a PROC MEANS statement is used to summarize the results of the simulation.

Flowchart for CALSIM Program



Simple Mediation Model (Appendix A)

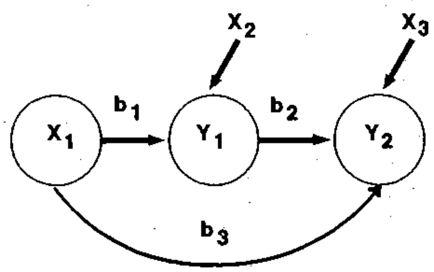
To illustrate how the program is used to evaluate the bias of analytical solutions, the standard error of the mediated effect is described in more detail here. The mediated effect describes the extent to which the relationship between two variables is due to an intervening or mediating variable. Mediated effects are particularly important in social and behavioral research because they refer to the underlying process of how variables are related to each other. In prevention research for example, mediated effects refer to the reason by which the prevention program had its effect on the outcome variable (MacKinnon et al., 1990). More on mediation can be found in James & Brett (1984), Baron & Kenny (1986), Judd & Kenny (1981), MacKinnon & Dwyer (1992), and Winship & Mare (1983).

The following diagram describes the simplest mediation model. The equations for Y_1 and Y_2 that define the model and the covariance matrix based on the equations are shown below. The equations are simulated in the DATA step and the parameters for the model are estimated using CALIS. The estimation procedure is repeated to determine the bias and sampling variability of several statistics related to the assessment of mediation.

The mediated effect in this model is defined as $b_1 b_2$ and its exact variance is equal to $\sigma^2_{b1} b^2_2 + \sigma^2_{b2} b^2_1 + \sigma^2_{b1} \sigma^2_{b2}$ (Goodman, 1960; Rice, 1988; Mood, Graybill, & Boes, 1974). Sobel (1982; 1986) proposed a general method for obtaining the variance of the mediated effect for this and other more complicated mediation models that reduces to $\sigma^2_{b1} b^2_2 + \sigma^2_{b2} b^2_1$ in this simple model. The exact variance formula is the second order Taylor series solution for the variance of the multiplication of two random variables. The standard error in Sobel (1982) is the first order Taylor series solution.

The simulation input values of sample size (NOBS) and parameter values are all that is needed to calculate the true standard error of b_1 and b_2 and the standard error of $b_1 b_2$. At each iteration, the discrepancy or bias from the true standard error and the simulated standard error can then be assessed. To be consistent with other literature (Baron & Kenny, 1986), parameters are also referred to as $a = b_1$, $b = b_2$, and $c = b_3$.

The purpose of this simulation was to assess the bias in the first and second order analytical solutions for the standard error of the mediated effect. In particular, the extent to which the bias in the estimators of the standard error of the mediated effect changed with sample size was investigated. The results of one such simulation are given in Table 1. Generally, as sample size increases, bias is reduced.



$$X_2,\ X_3 \sim N(0,1)$$

$$Y_1 = b_1 X_1 + X_2$$

$$Y_2 = b_2 Y_1 + b_3 X_1 + X_3$$

$$\begin{array}{c} \\ X_1 \\ Y_1 \\ Y_2 \end{array} \begin{array}{cc} X_1 & Y_1 \\ \left[\begin{array}{l} \sigma^2_{X1} \\ b_1 \sigma^2_{X1} \\ b_2 b_1 \sigma^2_{X1} + b_3 \sigma^2_{X1} \end{array}\right. & \left.\begin{array}{l} \\ b_1^2 \sigma^2_{X1} + \sigma^2_{X2} \\ b_1^2 b_2 \sigma^2_{X1} + b_1 b_3 \sigma^2_{X1} + b_2 \sigma^2_{X2} \end{array}\right] \end{array}$$

$$\begin{array}{cc} & Y_2 \\ Y_2 & \left[b_2^2 (b_1^2 \sigma^2_{X1} + \sigma^2_{X2}) + 2 b_2 b_2 b_1 \sigma_{X1}^2 + b_3^2 \sigma^2_{X1} + \sigma^2_{X3}\right] \end{array}$$

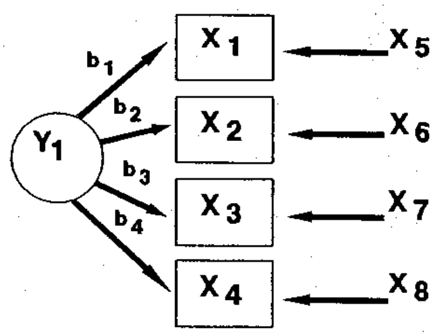
Table 1

Absolute Value of Bias

	Sample Size				
	25	50	100	200	1000
$b_1 b_2$	.0077	.0061	.0082	.0075	.0035
first	.0039	.0050	.0009	.0010	.0002
second	.0005	.0031	.0015	.0008	.0001

<u>One Factor Measurement Model (Appendix B)</u>

The mediation model described above can be easily estimated using PROC REG. Indeed, prior statistical simulations of this model were conducted using PROC REG (MacKinnon & Dwyer, 1990; MacKinnon, Warsi & Dwyer, 1991). It is not possible, however, to estimate measurement models using the REG procedure but the CALIS procedure is ideal for this. The following model is a measurement model for a single latent measure with four indicators. The program listed in Appendix B is used to simulate the one factor model.



$X_5, X_6, X_7, X_8 \sim N(0,1)$

$Y_1 = \sigma^2_{Y1}$

$X_1 = b_1 Y_1 + X_5$
$X_2 = b_2 Y_1 + X_6$
$X_3 = b_3 Y_1 + X_7$
$X_4 = b_4 Y_1 + X_8$

	X_1	X_2	X_3	X_4
X_1	$b_1^2\sigma^2_{Y1} + \sigma^2_{X5}$			
X_2	$b_2 b_1 \sigma^2_{Y1}$	$b_2^2\sigma^2_{Y1} + \sigma^2_{X6}$		
X_3	$b_3 b_1 \sigma^2_{Y1}$	$b_3 b_2 \sigma^2_{Y1}$	$b_3^2\sigma^2_{Y1} + \sigma^2_{X7}$	
X_4	$b_4 b_1 \sigma^2_{Y1}$	$b_4 b_2 \sigma^2_{Y1}$	$b_4 b_3 \sigma^2_{Y1}$	$b_4^2\sigma^2_{Y1} + \sigma^2_{X8}$

To illustrate this method we present the bias in each of the loadings (b_1, b_2, b_3, and b_4) as a function of sample size. Bias is reduced as sample size increases.

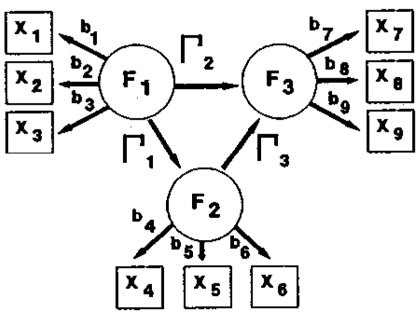
Table 2

Absolute Value of Bias

	Sample Size			
	50	100	200	1000
b_1	.0538	.0088	.0140	.0044
b_2	.0022	.0048	.0038	.0015
b_3	.0080	.0532	.0246	.0010
b_4	.0342	.0306	.0020	.0015

Three Factor Mediation Model (Appendix C)

The final model includes a mediation structural model along with measurement models for each of the latent constructs. The equations are presented below. The covariances are not presented because it would require too much space.



$F_2 = \Gamma_1 F_1 + D_1$
$F_3 = \Gamma_2 F_1 + \Gamma_3 F_2 + D_2$
$X_1 = b_1 F_1 + X_{10}$
$X_2 = b_2 F_1 + X_{11}$
$X_3 = b_3 F_1 + X_{12}$
$X_4 = b_4 F_2 + X_{13}$
$X_5 = b_5 F_2 + X_{14}$
$X_6 = b_6 F_2 + X_{15}$
$X_7 = b_7 F_3 + X_{16}$
$X_8 = b_8 F_3 + X_{17}$
$X_9 = b_9 F_3 + X_{18}$

A simulation study of the standard error of $\Gamma_1\Gamma_2$ was conducted. With a sample size of 1000, the first and second order Taylor series solutions equalled .0264 and the sampling variability of $\Gamma_1\Gamma_2$ across the 100 replications was .0257.

Conclusions

The CALSIM program can be used to conduct statistical simulations of covariance structure models. Several examples were illustrated. The program should make it easier for students and researchers to see how parameters change the covariance matrix, how sample size affects parameter estimates and standard errors, and the basis of covariance structure modeling in regression equations.

Future Directions

The program can be easily changed to simulate other covariance structure models. For example, complicated longitudinal models and mediational models with three parts have been simulated with this program. The effects of nonnormal distributions could be investigated by generating data using the IML program prior to using the simulation program (Chou, 1992). The MACRO and iterative features of the program could be used to bootstrap and jackknife estimates. It will also be helpful to obtain the true approximate standard errors for any model using the theoretical covariance matrix and the OMETHOD command in CALIS.

The purpose of this paper was to introduce statistical simulations in CALIS. More attention to efficiency would improve the program. The combination of the covariance structure modeling capabilities of the CALIS program along with the powerful capabilities of the other SAS procedures makes SAS an ideal environment for statistical simulations.

Author Note

This work was supported in part by Public Health Service grant #DA039116.

References

Baron, R.M. & Kenny, D.A. (1986). The moderator-mediator distinction in social psychological research:Conceptual, strategic, and statistical considerations. Journal of Personality and Social Psychology, 51, 1173-1182.

Bentler, P. M. (1989). Theory and Implementation of EQS: A Structural Relations Program. BMDP Statistical Software: Los Angeles, California, 1989.

Bentler, P.M. (1980). Multivariate analysis with latent variables: Causal modeling. Annual Review of Psychology, 31, 419-456.

Bollen, K.A. (1989). Structural Equations With Latent Variables. New York: Wiley.

Boomsma, A. (1985). Nonconvergence, improper solutions, and starting values in LISREL maximum likelihood estimation. Psychometrics, 50, 229-242.

Browne, M.W. (1982). Covariance structures. I.D. M. Hawkins (Ed.). Topics in Applied Multivariate Analysis. Cambridge, MA: Cambridge University Press.

Chou, C.P. (1992). A SAS/IML procedure to simulate data with nonnormal distribution. 1992 SAS User's Group International Meeting.

Goodman, L. A. (1960). On the exact variance of products. Journal of the American Statistical Association, 55, 708-713.

Hayduk, L.A. (1987). Structural Equation Modeling With LISREL: Essentials and Advances. Baltimore: The John Hopkins University Press.

James, L. R., & Brett, J. M. (1984). Mediators, moderators, and tests for mediation'. Journal of Applied Psychology, 69, 307-321.

Joreskog, K. G. & Sorbom, D. (1988). LISREL VII. Chicago: SPSS Inc.

Judd, C.M. & Kenny, D. A. (1981). Process analysis: Estimating mediation in treatment evaluations. Evaluation Review, 5, 602-619.

MacKinnon, D.P., Johnson, C.A., Pentz, M.A., Dwyer, J.H., Flay, B.R., Hansen, W.B., & Wang, E. (1990). Mediating mechanisms in a school-based drug prevention program: One year effects of the Midwestern Prevention Project", Health Psychology, 10(3), 164-172.

MacKinnon, D.P., Warsi, G., & Dwyer, J.H. (1991). A simulation study of the variance of indirect effect measures. Paper presented at the 1991 Psychometric Society Meeting.

MacKinnon, D.P., & Dwyer, J.H. (1992). Measurement of mediation in prevention studies. Manuscript submitted for publication.

McArdle, J.J. & Epstein, D. (1987). Latent growth curves with developmental structural equation models. Child Development, 58, 110-133.

Mood, A., Graybill, F.A., & Boes, D.C. (1974). Introduction to the Theory of Statistics. New York: McGraw-Hill.

Rice, J.A. (1988). Mathematical Statistics and Data Analysis. Pacific Grove, California, Wadsworth and Brooks.

Saris, W.E., & Stronkhorst, L.H. (1984). Causal modeling in nonexperimental research. Amsterdam: Sociometric Research Foundation.

Schoenberg, R. (1987). LINCS: Linear Covariance Structure Analysis Users Guide. RJS Software, Kensington, Maryland.

Sobel, M. E. (1982). Asymptotic confidence intervals for indirect effects in structural equation models. In S. Leinhardt (Ed.), Sociological Methodology 1982, (pp. 290-293). Washington, DC: American Sociological Association.

Sobel, M. E. (1986). Some new results on indirect effects and their standard errors in covariance structure models. In N. Tuma (Ed.), Sociological Methodology 1986, (pp. 159-186). Washington, DC: American Sociological Association.

Winship, C., & Mare, R.D. (1983). Structural equations and path analysis for discrete data. American Journal of Sociology, 89, 54-110.

APPENDIX A

```
OPTIONS PS=59 LS=80 REPLACE NONOTES;
%MACRO SIMULATE(NREP,NOBS,B1,B3,B2);
DATA SUMMARY; SET _NULL_;
%DO I=1 %TO &NREP;
TITLE 'SIMPLEST MEDIATION MODEL SIMULATION';
*GENERATE*;
DATA SIM;
DO I=1 TO &NOBS;
X1=RANNOR(0);
Y1=&B1*X1+RANNOR(0);
Y2=&B3*X1+&B2*Y1+RANNOR(0);
OUTPUT;
END;
*ESTIMATE*;
PROC CALIS DATA=SIM METHOD=ML NOPRINT COV OUTEST=OUT;
LINEQS
Y1=A  X1 + E1,
Y2=C  X1 + B Y1 +E2;
STD
E1=EE1,
E2=EE2;
DATA B; SET OUT;
IF _TYPE_='PARMS';
KEEP A B C;
DATA SE; SET OUT;
IF _TYPE_='STDERR';
SEA=A; SEB=B; SEC=C;
KEEP SEA SEB SEC;
DATA ALL; MERGE B SE;
DATA TEST;SET ALL;
NOBS=&NOBS;AB=A*B;
STRUEB1=SQRT(1/(&NOBS-2));
STRUEB2=SQRT(1/(&NOBS-3));
TRUESAB=SQRT(&B1*&B1*STRUEB2*STRUEB2+&B2*&B2*
STRUEB1*STRUEB1+STRUEB1*STRUEB1*STRUEB2*STRUEB2);
FIRST=SQRT(A*A*SEB*SEB+B*B*SEA*SEA);
SECOND=SQRT(B*B*SEA*SEA+A*A*SEB*SEB+SEA*SEA*SEB*SEB);
BAB=AB-(&B1*&B2);
BFIRST=FIRST-TRUESAB;
RFIRST=BFIRST/TRUESAB;
MSFIRST=BFIRST*BFIRST;
BSECOND=SECOND-TRUESAB;
RSECOND=BSECOND/TRUESAB;
MSECOND=BSECOND*BSECOND;
DATA NEW; SET SUMMARY;
DATA SUMMARY; SET NEW TEST;
%END;
PROC MEANS DATA=SUMMARY MEAN STD;
OUTPUT OUT=OUT;
%MEND;
PROC DATASETS LIB=WORK MT=DATA NOLIST;
DELETE SUMMARY SIM B SE OUT TEST ALL;
%SIMULATE(NREP=100,NOBS=50,B1=.5,B3=.2,B2=.6);
%SIMULATE(NREP=100,NOBS=100,B1=.5,B3=.2,B2=.6);
%SIMULATE(NREP=100,NOBS=200,B1=.5,B3=.2,B2=.6);
%SIMULATE(NREP=100,NOBS=1000,B1=.5,B3=.2,B2=.6);
```

APPENDIX B

```
OPTIONS PS=59 LS=80 REPLACE NONOTES;
%MACRO SIMULATE(NREP,NOBS,B1,B2,B3,B4);
DATA SUMMARY; SET _NULL_;
%DO I=1 %TO &NREP;
TITLE 'ONE FACTOR MEASUREMENT MODEL SIMULATION';
*GENERATE;
DATA SIM;
DO I=1 TO &NOBS;
F1=1*RANNOR(0);
X1=&B1*F1+1*RANNOR(0);
X2=&B2*F1+1*RANNOR(0);
X3=&B3*F1+1*RANNOR(0);
X4=&B4*F1+1*RANNOR(0);
OUTPUT;
END;
*ESTIMATE;
PROC CALIS DATA=SIM METHOD=ML COV NOPRINT OUTEST=OUT;
VAR X1 X2 X3 X4;
LINEQS
X1= B1 (&B1) F1 + E1,
X2= B2 (&B2) F1 + E2,
X3= B3 (&B3) F1 + E3,
X4= B4 (&B4) F1 + E4;
STD
E1=EE1,
E2=EE2,
E3=EE3,
E4=EE4,
F1=1.0;
DATA B; SET OUT;
IF _TYPE_ ='PARMS';
KEEP B1 B2 B3 B4 EE1 EE2 EE3 EE4;
DATA SE; SET OUT;
IF _TYPE_='STDERR';
SEB1=B1; SEB2=B2; SEB3=B3; SEB4=B4;
KEEP SEB1 SEB2 SEB3 SEB4;
DATA ALL; MERGE B SE;
DATA TEST; SET ALL;
BIASB1=B1-&B1;
BIASB2=B2-&B2;
BIASB3=B3-&B3;
BIASB4=B4-&B4;
DATA NEW; SET SUMMARY;
DATA SUMMARY; SET NEW TEST;
%END;
PROC MEANS DATA=SUMMARY N MEAN STD;
OUTPUT OUT=OUT;
%MEND;
PROC DATASETS LIB=WORK MT=DATA NOLIST;
DELETE B SIM OUT SE ALL TEST NEW SUMMARY;RUN;
%SIMULATE(NREP=100,NOBS=50,B1=-.4,B2=0.4,B3=-.8,B4=.6);
%SIMULATE(NREP=100,NOBS=100,B1=-.4,B2=0.4,B3=-.8,B4=.6);
%SIMULATE(NREP=100,NOBS=200,B1=-.4,B2=0.4,B3=-.8,B4=.6);
%SIMULATE(NREP=100,NOBS=1000,B1=-.4,B2=0.4,B3=-.8,B4=.6);
```

APPENDIX C

```
OPTIONS PS=59 LS=80 REPLACE NONOTES;
%MACRO SIMULATE(NREP,NOBS,B1,B2,B3,B4,B5,B6,B7,B8,B9,A,C,B);
DATA SUMMARY; SET _NULL_;
%DO I=1 %TO &NREP;
TITLE 'THREE FACTOR MEASUREMENT MODEL SIMULATION';
*GENERATE*;
DATA SIM;
DO I=1 TO &NOBS;
F1=RANNOR(0);
F2=&A*F1+RANNOR(0);
F3=&C*F1+&B*F2+RANNOR(0);
x1=&B1*F1+.00*RANNOR(0);
X2=&B2*F1+.51*RANNOR(0);
X3=&B3*F1+.51*RANNOR(0);
X4=&B4*F2+.00*RANNOR(0);
X5=&B5*F2+.51*RANNOR(0);
X6=&B6*F2+.51*RANNOR(0);
X7=&B7*F3+.00*RANNOR(0);
X8=&B8*F3+.51*RANNOR(0);
X9=&B9*F3+.51*RANNOR(0);
OUTPUT;
END;
*ESTIMATE*;
PROC CALIS DATA=SIM METHOD=ML COV CORR NOPRINT OUTEST=OUT;
VAR X1 X2 X3 X4 X5 X6 X7 X8 X9 ;
LINEQS
X1=  .7  F1 + E1,
X2= B2 (&B2) F1 + E2,
X3= B3 (&B3) F1 + E3,
X4=  .7  F2 + E4,
X5= B5 (&B5) F2 + E5,
X6= B6 (&B6) F2 + E6,
X7=  .7  F3 + E7,
X8= B8 (&B8) F3 + E8,
X9= B9 (&B9) F3 + E9,
F2=A (&A)  F1 + D2,
F3=C (&C)  F1 + B (&B) F2 +D3;
STD
F1=DD1 (1),
D2=DD2 (1),
D3=DD3 (1),
E1=EE1  0,
E2=EE2 (.51),
E3=EE3 (.51),
E4=EE4 0,
E5=EE5 (.51),
E6=EE6 (.51),
E7=EE7 0,
E8=EE8 (.51),
E9=EE9 (.51);
DATA B; SET OUT;
IF _TYPE_='PARMS';
KEEP A B C;
DATA SE; SET OUT;
IF _TYPE_='STDERR';
SEA=A; SEB=B; SEC=C;
KEEP SEA SEB SEC;
DATA ALL; MERGE B SE;
DATA TEST;SET ALL;
*STANDARD ERROR OF THE MEDIATED EFFECT;
FIRST=SQRT(A*A*SEB*SEB+B*B*SEA*SEA);
SECOND=SQRT(B*B*SEA*SEA+A*A*SEB*SEB+SEA*SEA*SEB*SEB);
AB=A*B;
BAB=AB-(&A*&B);
DATA NEW; SET SUMMARY;
DATA SUMMARY; SET NEW TEST;
%END;
PROC MEANS DATA=SUMMARY MEAN STD;
OUTPUT OUT=OUT;
%MEND;
PROC DATASETS LIB=WORK MT=DATA NOLIST;
DELETE SIM SUMMARY B SE ALL TEST OUT;
%SIMULATE(NREP=100,NOBS=50,B1=0.7,B2=0.7,B3=.7,B4=.7,b5=.7,
B6=0.7,B7=0.7,B8=0.7,B9=.7,A=.5,C=.2,B=.6);
%SIMULATE(NREP=100,NOBS=100,B1=0.7,B2=0.7,B3=.7,B4=.7,b5=.7,
B6=0.7,B7=0.7,B8=0.7,B9=.7,A=.5,C=.2,B=.6);
%SIMULATE(NREP=100,NOBS=200,B1=0.7,B2=0.7,B3=.7,B4=.7,b5=.7,
B6=0.7,B7=0.7,B8=0.7,B9=.7,A=.5,C=.2,B=.6);
%SIMULATE(NREP=100,NOBS=1000,B1=0.7,B2=0.7,B3=.7,B4=.7,b5=.7,
B6=0.7,B7=0.7,B8=0.7,B9=.7,A=.5,C=.2,B=.6);
```