Data Analysis with Structural Equations[1]

The method of structural equations that is described in the PowerPoint file in Menu 1[2] can be executed on standard statistical packages such as the *Statistical Package for the Social Sciences* (SPSS). A brief instructional guide for accomplishing those analyses is presented here. Instructions are specified in terms of statements that can be used in a mainframe SPSS program. The same commands can be used in the PC versions with only minor modifications; the mainframe syntax is easier to explain, however.

Note the use of the term "structural equations" is consistent with the mathematical definition thereof. The procedures listed below do not involve Linear Structural Relations Analysis (LISREL) or its later versions that allow some types of nonlinear analysis.

CATASTROPHE MODELS WITH POLYNOMIAL REGRESSION

The data set for catastrophe models needs to be organized so that each observation contains the dependent measure (the variable that is hypothesized to show catastrophic behavior) at two points in time, and the values of the control value at Time 1. The user then needs to specify some COMPUTE statements before specifying the actual regression sub-program. COMPUTE statements are needed to transform dependent measures and other control variables with respect to location and scale.

```
COMPUTE z2 = (y2 – L)/ S                                          (A1)
```

In statement A1, *y2* is the dependent measure at Time 2, *L* is the value of location, and *S* is the value of scale. Actual numbers would replace *L* and *S*. The same syntax would be used for control variables and for changing $y_1 \rightarrow z_1$. Compute statements are also needed to define power terms such as z^3 (statement A2), a similar quadratic term, bifurcation interactive terms (A3), and the difference score (A4) that is used as the dependent measure in the catastrophe models.

```
COMPUTE zpow3 = z1**3                                             (A2)

COMPUTE bz = b*z1                                                 (A3)

COMPUTE deltaz = z2 – z1                                          (A4)
```

The variable *b* listed in statement A3 is a variable that is hypothesized to function as a bifurcation term in a cusp. A particular application can have several variables called *b*. The variable called *a* in statement A5 is a variable that is hypothesized to function as an asymmetry variable.

Proceed to the main program statements after completing the COMPUTE statements. Use the regression sub-program.

```
Regression descriptives/ missing=pairwise                         (A5)
      /variables = dz zpow3 zpow2 bz a
      /dependent = dz/ enter zpow3 zpow2 bz a
```

[1] Portions of this text are adapted from: Guastello (2002 Appendix A).
[2] "Resources for students and teachers," www.societyforchaostheory.org/tutorials

Next inspect the significance tests for each of the terms in the model. If there the regression weight for `zpow3` is not significant, drop `zpow2` and try it again. In the event that there are several variables being tested as *a* and *b* variables, drop any variable for which the *p*-value on the regression weight is greater than 0.10. Then try the reverse hypothesis, that the variables first thought to behave as *b* are really *a*'s and vice versa. To do so define some more compute statements for the new bifurcation terms (A6) and run the regression as a two-step process (A7).

```
COMPUTE az = a*z1                                                          (A6)
```

```
REGRESSION descriptives/ missing=pairwise                                  (A7)
      /variables = dz zpow3 zpow2 bz a b az
      /dependent = dz /enter zpow3 zpow2 a  /enter az b
```

Statement A7 assumes that the variable that was first thought to behave as *b* was not significant, and that both `zpow3` and `zpow2` were acceptable:

Finally, the two linear alternative models would be defined as A8 and A9:

```
REGRESSION  descriptives/ missing=pairwise                                 (A8)
      /variables = dz b a
      /dependent = dz /enter a b
```

```
REGRESSION descriptives/ missing=pairwise                                  (A9)
      /variables = z2 z1 b a
      /dependent = z2 /enter z1 b a
```

EXPONENTIAL SERIES WITH NONLINEAR REGRESSION

Nonlinear regression offers much greater flexibility in the definition of models compared to polynomial regression and its variations on the GLM. Because of its flexibility in defining a model, the nonlinear regression sub-program of SPSS requires a more specific statement of the intended models, and there is little automatic processing with regard to putting variables in or out of the model.

The data set up requires that the observations be ordered in time series, where each subsequent string of data represents observations at successive points in time. The control program again requires COMPUTE statements, but not generally as many. The definitions of *y* → z go in first along with the conversions for any control variables. Define the *z* as z_2 (statement A10) then use the LAG syntax to define z_1 at a lag of 1 time period (A11).

```
COMPUTE z2 = (y – L) / S                                                   (A10)
COMPUTE z1 = lag(z2, 1)                                                    (A11)
```

Next we define the three program statements for nonlinear regression. All three begin in the left-most space of the line; there is no indentation on the second or third command. The first line specifies the variables that will appear in the model as nonlinear regression weights. Nonlinear regression is an iterative calculation process whereby the user specifies some initialized values of the regression weights (*a*, *b*, and *c* in this example). The program then fits the initial values with the model function to the data, makes an adjustment, then fit the result to the derivative of the function, makes an adjustment, then re-fits the principal model to the data again, and so forth, until the resulting corrections become trivial.

```
MODEL PROGRAM a = 0.5, b = 0.5, c = 0.5                                    (A12)
```

It is usually good to specify initial values of the regression weights that are close to the final values, if the final values are known. I usually use 0.5 for initial values of all parameters in these studies. I have experimented (or rather fiddled) with possible strategies for using different initial values, but I have not yet found any strategy better than the equal estimates where nonlinear dynamics are concerned.

The second specifies the nonlinear model,

```
COMPUTE PRED = a*exp(b*z1) + c
```

(A13)

The third specifies the dependent measure and executes the program,

```
NLR z2 with z1
```

(A14)

The output from an example analysis corresponding to A13 appears in the last two pages of this document. For a recent exposition of the exponential modeling technique and an elaborate example involving two order parameters, see Guastello, Nathan, & M. Johnson (2009). For a verification of how well the technique performs in determining the fractal dimension of classic chaotic attractors, see T. Johnson and Dooley (1996).

Nonlinear regression programs sometime offer options such as constrained nonlinear regression or different methods of specifying error terms. Constrained nonlinear regression keeps the parameter estimates within certain boundary values, which are typically chosen based on previous studies of similar functions with similar data. In the absence of any good reason to constrain values one way or another, I recommend using the unconstrained nonlinear regression, which is specified in statement A14 by the command NLR.

Another option allows the user to select of the principle of maximum likelihood for calculating the error component of the regression model instead of the principle of least squares.

Contrary to what some people seem to think, however, there is no particular association between maximum likelihood and nonlinear modeling and least squares with linear modeling; both can be used in either type of regression, although it's true that linear regression with maximum likelihood is much less common than the least squares method.

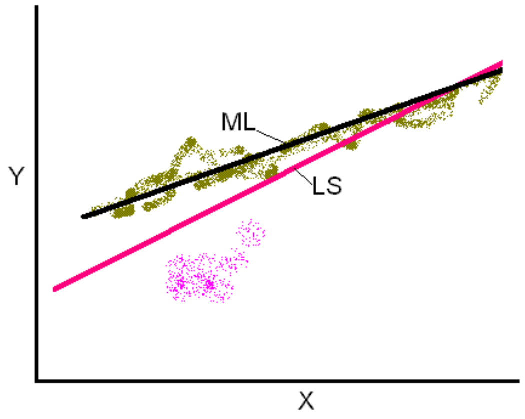


Fig. 1. The methods of least squares and maximum likelihood would place the line differently in the face of messy (pink) data.

Figure 1 illustrates how the two procedures would behave differently if they were looking for a line in messy data. The method of least squares, which is based on the principle of minimizing the squared distance from the regression line, would place the line in the location shown as LS. The method of maximum likelihood, however, would fit the line according to locations of greatest density, and thus place the line differently. Least squares would recognize the presence of the anomalous data points shown in pink, while maximum likelihood would essentially ignore them. For messy data, maximum likelihood might indeed have a greater chance of finding a good line where other events are occurring, but it does capitalize on chance. Least squares might have difficulty finding the line, but the line it would find would be more apt to generalize to situations where "pink data" are likely to occur again. For this reason some nonlinear analysts prefer the least squares technique. I would make the same recommendation; if you can find what you are looking for using least squares, the prognosis for generalizing out the original sample is greater and there is not much need to use alternative methods.

The nonlinear regression analysis concludes with an ANOVA table and a value of R^2 for the model overall. Tests on the specific regression weights are listed by SPSS as estimated values and their confidence intervals at the 95% level. If the upper and lower boundaries of the confidence interval are both positive or both negative for a particular regression weight, then the confidence does not include 0.00, and the results can be interpreted as significant at $p < .05$.

The weight on the exponent is the critical element of the analysis. In the event that statistical significance is not obtained for that weight, proceed to delete the less essential components from the models, namely, the constants. Statements A12 and A13 change to A15 and A16, respectively, while A14 remains the same:

```
MODEL PROGRAM b = 0.5                                                  (A15)
COMPUTE PRED = exp(b*z1)                                               (A16)
NLR z2 with z1                                                         (A14)
```

To test the bifurcation model, Statements A12 and A14 remain the same, and A13 becomes A15.

```
COMPUTE PRED = a*z1*exp(b*z1) + c                                      (A15)
```

Finally, there are times in which a nonlinear regression model fits so poorly that a negative R^2 is produced. Those values should be interpreted as equivalent in meaning as .00, meaning that the model fit was extremely poor.

CATASTROPHE PDFS WITH NONLINEAR REGRESSION

The method of testing catastrophe models by analyzing probability density functions (pdfs) through nonlinear regression is preferable, or should I say inevitable in two types of conditions: (a) The data are only measured at one point in time, but a catastrophe model is thought to exist therein nonetheless (e.g. in survey data, Smerz & Guastello, 2008), or (b) all time-1 measurements are 0.00 (e.g. in leadership emergence studies, Guastello & Bond, 2007).

Catastrophe models can be tested at two levels through the nonlinear regression procedure (a) the pdf only with no hypothesis about the control variables, and (b) with hypotheses about the control variables.

PDFs Without Control Variables

The test for the catastrophe pdfs is not substantially different from the exponential regression models just considered. There are two additional steps in data preparation, however. The first step requires a frequency distribution on the raw scores of the dependent measure, *y* that will produce the cumulative probability (or percentile) of *y*. Second, use a RECODE command to substitute the cumulative probabilities of *y* for *y* and give the result a new variable name, PCTY (A16).

```
RECODE Y (0 = .500) (1 = .617) (2 = .683) (3 = .725)                (A16)
      (4 = .750) (5 = .775) (6 = .800) (7 = .825) (8 = .842)
      (9 = .842) (10 = .883) (11 = .933) (12 = .950) (13 = .967)
      (14 = .999) (15 = .999) (16 = .999) into PCTY
```

The decimal values shown in A16 were those that were actually used in the leadership emergence study by Zaror and Guastello (2000) in which we were looking for a swallowtail catastrophe pdf. We still need to convert *y* to *z*. SPSS sometimes encounters an computational overflow when running nonlinear regression if, on one of the iterations, the numerical argument to the exponent exceeds 88. The problem is solved by multiplying *S* by 100 as shown in A17.

```
 COMPUTE z = (Y - L)/(S * 100)                                            (A17)
```

We can now proceed to the three statements for the nonlinear regression model. A19 is testing for the shape of the pdf only, and does not include the control variables. The control variables are essentially treated as constants and they are absorbed into the regression weights designated below as *a, b, c, d*, and *e*. Adaptations for testing the control variables are described later on.

```
MODEL PROGRAM  x= 0.5, a= 0.5 b=0.5 c=0.5 d=0.5 e=0.5               (A18)
COMPUTE PRED = x*exp(a*(z**5)+b*(z**4)+c*(z**3)+d*(z**2)+e*z)       (A19)
NLR PCTY with z                                                     (A20)
```

In the event that statistical significance is not obtained for each of the regression weights, the least essential element in the model can be dropped. In this case we would drop “`b=0.5`” from A18, and “`+b*(z**4)`” from A19; A20 would not change.

To test a cusp model instead of a swallowtail, substitute A21for A19:

```
COMPUTE PRED = x*exp(a*(z**4)+b*(z**3)+c*(z**2)+d*z)                (A21)
```

Determining Critical Points

Sometimes one is interested in the critical points associated with the catastrophe models – the values of the equilibria (attractor modes) and the repellor (statistical antimode). The critical values can be solved analytically by using the estimated parameters for *a, b, c, d,* and *e,* taking the second derivative of the argument to the exponent, setting it equal to 0.00, and solving for the roots. A regression program can do the job much more easily, however, than solving a differential equation by hand. To begin, create a data set that contains two columns: *Y* and the Frequency of *Y*, based on the frequency distribution was the previously obtained.

The analysis is a polynomial regression of Frequency of *y* as a function of *y*. The PSI-PLOT program (from Poly Software International) can perform the analysis instantly with a point-and-click to polynomial regression. The job can be done through SPSS, nonetheless. In either case, the conversion of *y* to *z* is not necessary for this type of problem. First compute polynomials of y^2, y^3, and y^4 using the syntax given in Statement A2. Then define the regression program to enter the polynomials from lowest to highest.

```
REGRESSION descriptives /missing = pairwise                         (A22)
      /variables = FREQY, y, ypow2, ypow3, ypow4
      /dependent = FREQY
      /enter y /enter ypow2 /enter ypow3 /enter ypow4
```

For data such as the leadership emergence data, there were only 17 values of *y*, so it was no surprise that all power polynomials showed statistical significance. The objective here is point estimation, however, rather than determination of significance. The regression results give regression weights that can be used to compute a predicted frequency of *y*. A plot of the predicted frequency of *y* gives the underlying modes and antimodes for the frequency distribution.

Testing Control Variables in Catastrophe Models

The early means of testing control variables in catastrophe models with nonlinear regression (Cobb, 1981) was indirect. The procedure first determined the cusp pdf, then produced parameter estimates for all observations. Then, as a separate step, it would calculate linear correlations between the hypothesized control variables and the parameter estimates associated with bifurcation and asymmetry. The matrix of correlations can be interpreted in much the same way as one would interpret the results of a factor analysis, where some research variables would "load" on one control parameter, and other variables would load on the other.

One limitation of the indirect strategy is that the R^2 for the model would only pertain to the pdf itself, and not include the impact of the control variables. Control variables, if they were placed in the model would, incidentally, have the effect of lowering the overall R^2 a bit because they are only imperfectly related to the parameter estimates. Another limitation is that the indirect method does not allow for point estimation except perhaps in a clumsy manner.

Thus we move on to a direct method for testing control variables as part of the catastrophe model. It helps enormously to have a clear hypothesis for the entire catastrophe model, and not simply the shape of the model. If the research situation lends itself to many possible variables that could be control variables, it is advisable to factor analyze them first, and then work with the common factors. For an example where factor analysis helped the cause, see Guastello and Bond (2007).

The nonlinear regression procedure involves only a small adaptation to what was presented already. For a cusp, let V_1 and V_2 represent hypothesized control variables for asymmetry and bifurcation respectively. A21 then becomes:

COMPUTE PRED = x*exp(a*(z**4)+b*(z**3)+c*V_2*(z**2)+d*V_1*z) (A23)

For a swallowtail model, we would also have a control variable V_3 for bias. Thus A19 becomes:

COMPUTE PRED = x*exp(a*(z**5)+b*(z**4)+c*V_3*(z**3)+d*V_2*(z**2)+e*V_1*z) (A24)

In A23 and A24, the SPSS variable name would be substituted where V_1 V_2 and V_3 are shown.

REFERENCES

Guastello, S. J. (2002). *Managing Emergent Phenomena*. Mahwah, NJ: Lawrence Erlbaum Associates.

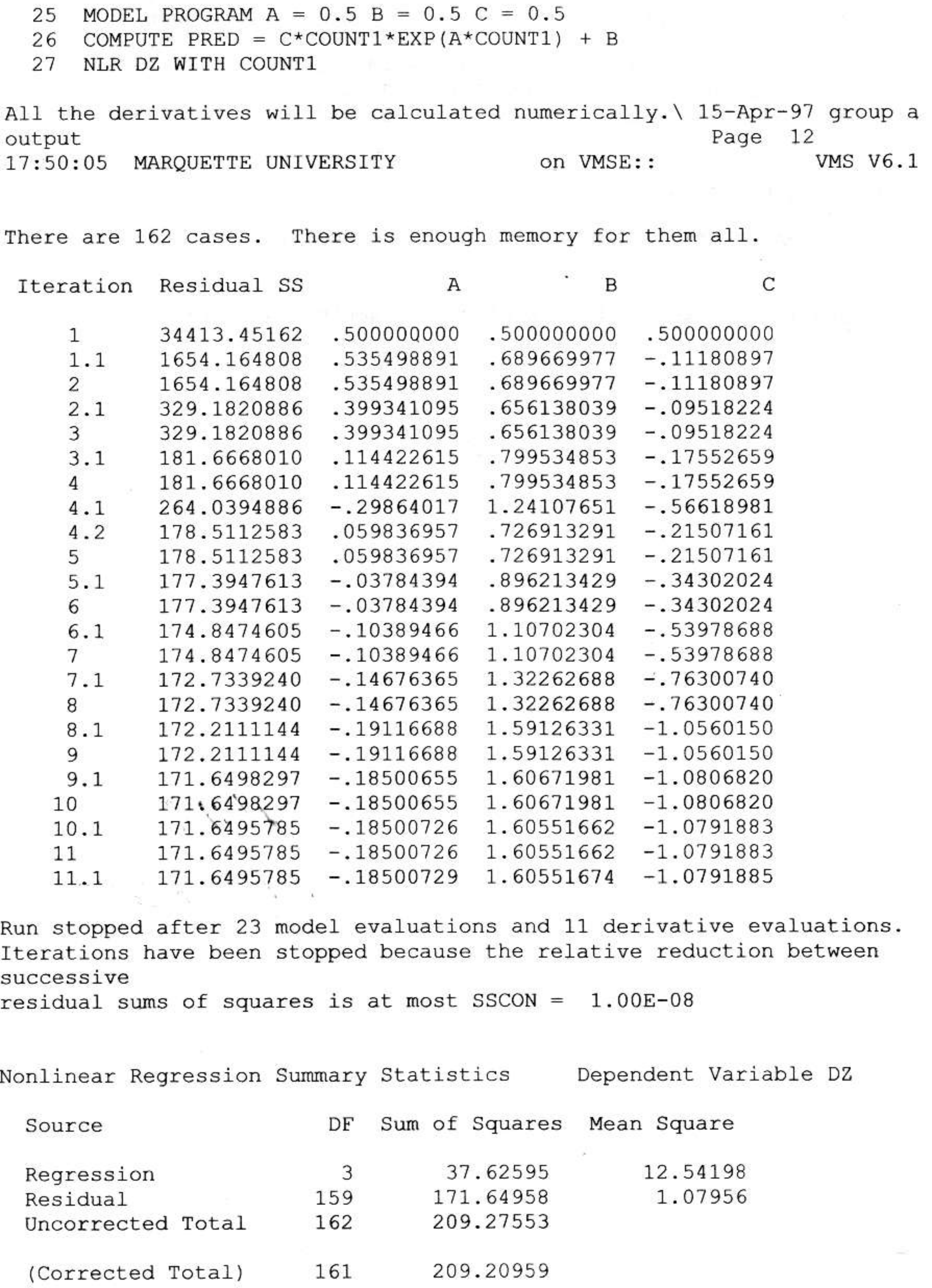
Guastello, S. J., & Bond, R. W. Jr. (2007). A swallowtail catastrophe model of leadership in coordination-intensive groups. *Nonlinear Dynamics, Psychology, and Life Sciences, 11, 235-351.*

Guastello, S. J., Nathan, D. E., & Johnson, M. J. (2009). Attractor and Lyapunov models for reach and grasp movementions with application to robot-assisted therapy. *Nonlinear Dynamics, Psychology, and Life Sciences, 13,* in press.

Johnson, T. L., & Dooley, K. J. (1996). Looking for chaos in time series data. In W. Sulis & A. Combs (Eds.), *Nonlinear dynamics in human behavior* (pp. 44-76). Singapore: World Scientific.

Smerz, K. E., & Guastello, S. J. (2008). Cusp catastrophe model for binge drinking in a college population. *Nonlinear Dynamics, Psychology, and Life Sciences, 12, 205-224.*

Zaror, G., & Guastello, S. J. (2000). Self-organization and leadership emergence: A cross-cultural replication. *Nonlinear Dynamics, Psychology, and Life Sciences, 4,* 113-120.



```
  25  MODEL PROGRAM A = 0.5 B = 0.5 C = 0.5
  26  COMPUTE PRED = C*COUNT1*EXP(A*COUNT1) + B
  27  NLR DZ WITH COUNT1

All the derivatives will be calculated numerically.\ 15-Apr-97 group a
output                                                     Page  12
17:50:05  MARQUETTE UNIVERSITY            on VMSE::               VMS V6.1


There are 162 cases.  There is enough memory for them all.

 Iteration  Residual SS           A           B           C

     1       34413.45162  .500000000  .500000000  .500000000
     1.1     1654.164808  .535498891  .689669977  -.11180897
     2       1654.164808  .535498891  .689669977  -.11180897
     2.1     329.1820886  .399341095  .656138039  -.09518224
     3       329.1820886  .399341095  .656138039  -.09518224
     3.1     181.6668010  .114422615  .799534853  -.17552659
     4       181.6668010  .114422615  .799534853  -.17552659
     4.1     264.0394886  -.29864017  1.24107651  -.56618981
     4.2     178.5112583  .059836957  .726913291  -.21507161
     5       178.5112583  .059836957  .726913291  -.21507161
     5.1     177.3947613  -.03784394  .896213429  -.34302024
     6       177.3947613  -.03784394  .896213429  -.34302024
     6.1     174.8474605  -.10389466  1.10702304  -.53978688
     7       174.8474605  -.10389466  1.10702304  -.53978688
     7.1     172.7339240  -.14676365  1.32262688  -.76300740
     8       172.7339240  -.14676365  1.32262688  -.76300740
     8.1     172.2111144  -.19116688  1.59126331  -1.0560150
     9       172.2111144  -.19116688  1.59126331  -1.0560150
     9.1     171.6498297  -.18500655  1.60671981  -1.0806820
    10       171.6498297  -.18500655  1.60671981  -1.0806820
    10.1     171.6495785  -.18500726  1.60551662  -1.0791883
    11       171.6495785  -.18500726  1.60551662  -1.0791883
    11.1     171.6495785  -.18500729  1.60551674  -1.0791885

Run stopped after 23 model evaluations and 11 derivative evaluations.
Iterations have been stopped because the relative reduction between
successive
residual sums of squares is at most SSCON =  1.00E-08


Nonlinear Regression Summary Statistics     Dependent Variable DZ

  Source                  DF  Sum of Squares  Mean Square

  Regression               3        37.62595      12.54198
  Residual               159       171.64958       1.07956
  Uncorrected Total      162       209.27553

  (Corrected Total)      161       209.20959
  R squared = 1 - [Residual SS / Corrected SS] =      .17953

                                                  Asymptotic 95 %
                                  Asymptotic    Confidence Interval
  Parameter     Estimate        Std. Error     Lower          Upper

  A           -.185007292    .056955741  -.297494665  -.072519919
  B           1.605516744    .403831106   .807951839  2.403081650
  C          -1.079188522    .422936783 -1.914487067  -.243889978


15-Apr-97 group a output
Page  13
17:50:07  MARQUETTE UNIVERSITY             on VMSE::               VMS V6.1

  Asymptotic Correlation Matrix of the Parameter Estimates

                  A          B          C

  A            1.0000     -.7024      .9016
  B            -.7024     1.0000     -.9285
  C             .9016     -.9285     1.0000
```