SHAM-REG 2

SHAMMAS SOFTWARE SERVICES
1533F HONEY GROVE DRIVE
RICHMOND, VIRGINIA 23229

DISCLAIMER OF ALL WARRANTY AND LIABILITY

The author of the program makes no warranties, either expressed of implied, with respect to this manual or with respect to the software described in this manual, its quality, performance, merchantability or fitness for any particular purpose. SHAM-REG 2 is sold "As Is". While every effort is made to make SHAM-REG 2 versatile, in no event will Shammas Software Services and/or the author be liable for direct, indirect, incidental or consequential damages resulting from any defect in the software.

TABLE OF CONTENTS

FORWARD

SHAM-REG 2 is a multiple linear regression pac that will run on the Hewlett-Packard HP 41CV .

The additional hardware needed are :

(1) The HP-IL interface loop module (#82160A)

(2) The digital cassette drive (#82161A)

(3) The Extended functions module (#82180A)

(4) One extended memory module (#82181A).

The use of DATA BASE 1 is recommended to edit and manipulate stored data.

All of the hardware products mentioned above are manufactured by the Hewlett-Packard Co.

DATA BASE1 is a product of SHAMMAS Software Services.

INTRODUCTION

It is important for the user to share some of the views regarding the types of regression with which this package was planned and written.

We can say that regression analysis is used to explain the variations between two or more variables by assuming that a proposed equation (model), derived analytically or empirically, is adequate (at least for the user's purposes) for such correlation. The proposed model should state a relationship between the dependent variable and one or more independent variable. Constants and possibly suitable transformations will also be involved in the equations, and it is these constants that regression seeks to evaluate and give a sense of measured effect as well as to provide the user with an equation that can be used to predict values for unavailable data.

It is because of the presence of mathematical functions (transformations) that regression analysis is not always able to work in a simple straight-forward fashion. Indeed, it is how we obtain the regression constants (i.e. coefficients) that determines the category of the model we are using. There are two global categories ;

(1) General Linear Regression : This is characterized by the fact

that all the variables involved will NOT be transformed in any way ; instead they are processed "as they are" . This category is divided into two classes ;

(1.1) Linear Regression : This is "the" simplest form where the dependent variable, call it Y, is affected by one independent variable, call it X, in the following manner,

$$Y = a + b\ X \qquad (1)$$

where a= intercept, b=slope.

"a" and "b" are the regression coefficients. The reader will most probably recognize this class, beacuse of its popular implementation in pocket calculators.

(1.2) Multiple Linear Regression : Similar to the above, we have here an extention in the number of independent variables. An example is,

$$Y = a + b\ X1 + c\ X2 \qquad (2)$$

$$Y = a + b\ X1 + c\ X2 + d\ X3 \qquad (3)$$

where X1,X2 and X3 are the independent variables and "a","b","c" and "d" are the regression coefficients.

In this category the regression coefficients are obtained by carrying out the proper summations and then solving the simultaneous linear equations yielding the desired results.

(2) General Nonlinear Regression : Due to the use of functions in

describing the relationship between the variables, we can no longer proceed as we did in the first category. The steps taken to overcome this problem will cause two classes to be distinguished into ;

(2.1) Intrinsically Linear : Suitable transformations on our variables may linearize the model used and make the situation look like a category 1 regression. Examples are ;

Power curve : $Y = a X^b$ (4)

Transformations using the logarithms on both sides of the equation gives,

$\log(Y) = \log(a) + b \log(X)$ (5)

Empirical curve : $Y = a X1^(b) X2^(c)$ (6)

Again a logarithmic transformation will give,

$\log(Y) = \log(a) + b \log(X1) + c \log(X2)$ (7)

Transforming equations (4) and (6) gave us equations (5) and (7) that possess linearity and thus are similar to equations (1) and (2). The solution will consist of ;

(1) Transformation of the variables.

(2) Summing up the proper terms.

(3) Solving the simultaneous equations.

(4) Possible "inverse transformations" of the regression coefficients obtained.

In item (4) above we note that because of the transformations some or all the regression coefficients obtained, will also be

transformed and one may need to carry out the inverse transformation to obtain the values to be used in the "original" model. In equation (5) and (7) the coefficient "a" will be obtained as log(a).

(2.2) Intrinsically Nonlinear : Due to the nature of the model, the steps taken in the first class of this category may not work (i.e no suitable transformation can be used to "linearize" the model). Iterative processes are involved and would work as follow :

(1) Obtain an initial guess for the coefficient(s) preventing linearization.

(2) With the coefficient(s) above assumed, we now linearize the model.

(3) Carry out the summations.

(4) Solve the simultaneous equations.

(5) Check for optimum sum of least square or use any other criteria to determine convergence. The latter is affected by the values of the assumed coefficient(s).

(6) If the results are satisfactory and calculations, else improve the guesses and go back to step (2).

An example is the crescent shaped curve,

$$Y = A\,[\,1 - \exp(-B\,X)\,] \qquad (8)$$

We have two possibilities,

(1) Assume a value for "A" and transform to

$\ln(A - Y) = \ln(A) - B X$

the "A" on the left is the assumed value, while the one to the right is calculated. We minimize the sum of squares by changing "A".

(2) We can linearize using the Taylor series, assume a value for "B" and minimize the sum of squared errors by iteration. In a few cases the model may be transformed in such a way that iteration is avoided if the dependent variable is obtained in a regular pattern, such as equal intervals.

The two categories presented are extreme cases. A "hybrid" combination of category 1 (class 2) and category 2 (class 1) exists. It is the case when :

(1) Variable transformation may not be applied to all the variables involved as is ;

$Y = a + b X1 + c \log(X2) + d (1/X3)$ (9)

where Y and X1 remain untransformed.

(2) Variables may co-exist with their transformations side by side. This means having more terms in the model than the total number of variables. An example is the class of powerful polynomials,

$Y = a + b X + c X^2 + d X^3$ (10)

the above is basically a model to relate Y and X, where the square and cube of X (its transformation) form additional terms

in the model. Another example is,

$$Y = a + b\ X1 + c\ X1^2 + c\ \log(X3) + d\ X3 \qquad (11)$$

where more than one independent variable show up. It is a four term model that relates the dependent variable with three independent variables.

SHAM-REG 2

This program package will perform multiple linear regression with the following features :

(1) Input/Output : The data is read from the cassette tape. Keyboard entry is allowed but only to store the data on mass media. It is worthwhile to point out that as far as the stored data are concerned the program has no preconceived notion on which variable is the dependent variable and which are the independent variables and consequently the stored data variables will be called "input" variables. To manipulate and edit the data the user will need a data base program such as DATA BASE 1. Up to ten variables can be inputed per data point.

(2) Regression : One of the main features of SHAM-REG 2 is that the user can select and screen the input variable while informing the program about the model for the regression. For example the data consist of input variables X1 to X8. The user may tell the program that he wishes to carry out a multiple regression with three independent variable terms using the following model :

$$1/X5 = b0 + b1\ LnX2 + b2\ X2 + b3\ X3$$

thus regression variable #1 = X2 (with a log transformation)

regression variable #2 = X2

regression variable #3 = X3

dependent variable = X5 (with a reciprocal transformation). As one can see only X2, X3 and X5 were selected with X2 occuring in two terms and the rest of the variables were not considered.

THE RULES FOR VARIABLES SELECTION ARE:

(1) AT LEAST TWO INPUT VARIABLES SHOULD ENTER THE REGRESSION OF UP TO 10 TERMS.

(2) THE INPUT VARIABLE SELECTED AS THE DEPENDENT VARIABLE MAY OCCUR ON BOTH SIDES OF THE EQUAL SIGN IN THE MODEL.

(3) NO TWO TERMS MAY CONTAIN THE SAME VARIABLE WITH THE SAME TRANSFORMATION.

(4) THE VARIABLES DURING TRANSFORMATIONS ARE EXPECTED TO GIVE VALID RESULTS. THUS NO NEGATIVE VALUES ARE ALLOWED WITH LOGARITHMIC TRANSFORMATIONS...ETC.

A word about transformations. Most of the available programs use routines that take the values of the dependent variable and independent variables, transform them IN THE SAME ROUTINE to the desired forms. This means that for 'n' linearized models there should be 'n' transformation routines. Not so with SHAM-REG 2 ! We have chosen to carry out these transformations such that each variable's transformation is done separately and the ensemble will form the requested model, as was explained in the tutorial part earlier. Thus over 8x10^11 models can be fitted using SHAM-REG 2 !

(3) Regression results : The programs will calculate the

regression coefficients, coefficient of determination, R^2, the ANOVA (Analysis of variance) table, standard errors for the regression coefficients and their corresponding critical student-t statistic values. Projections are available and so are the confidence intervals for the regression coefficients via a routine to evaluate the student-t statistic and spare the user the use of tables.

The programs use the correlation matrix to minimize the roundoff errors and also provide the inverse matrix. The latter is needed to calculate the standard errors. . These calculations will be slow because the program is doing a lot of calculations !

USER'S INSTRUCTIONS

During the following insructions a "R/S*" will indicate to skip the "R/S" if a printer is used.

(1) Clear the contents of the extended memory including the additional memory in the extended function module. This step is not needed if only "TS" and "MTS" data files, created by SHAM-REG 2, exist from previous runs.

(1.1) FUNCTION : XEQ EMDIR

DISPLAY : XMEMORY files.

(1.2) INPUT : [ALPHA] filename [ALPHA]

FUNCTION : XEQ PURFL

(1.3) Repeat step (1.2) to purge all files.

(2) Set memory partition "SIZE" and display format. This is only to insure that a reasonable space is available for programs to be read. The latter will adjust the "SIZE".

FUNCTION : SIZE nnn

FUNCTION : FIX n or ENG m or SCI 1

(3) Load the transformation functions library, if not in the memory already.

INPUT : [ALPHA] LIB2 [ALPHA]

FUNCTION : XEQ READP

FUNCTION : GTO . .

(4) (Optional) To input data from the keyboard and store on tape.

(4.1) Load program "R2I".

INPUT : [ALPHA] R2I [ALPHA]

FUNCTION : XEQ READP

(4.2) Run the loaded program.

FUNCTION : XEQ R2I

DISPLAY : FILENAME?

INPUT : name of data file under which the new data will be saved.

FUNCTION : R/S

(4.3) DISPLAY : STATUS?

INPUT : (Optional) either enter the name of an already existing status file or make no input and allow the program to create a new status file under the name "Sfilename".

FUNCTION : R/S

(4.4) DISPLAY : MAX DATA?

INPUT : Maximum number of data points.

FUNCTION : R/S

(4.5) DISPLAY : NVAR?

INPUT : # of variables per data point

FUNCTION : R/S

(4.6) A loop will start to prompt the input of all the variables (X1...Xm) for one data point.

DISPLAY : Xi?

INPUT : Xi

FUNCTION : R/S

(4.7) Repeat step (4.6) for all your data.

FUNCTION : R/S

Goto step (4.6)

Note: If you fill your data file then the program will display a message concerning the matter.

DISPLAY : FILE FULL

(4.8) If the data file is not full by the end of the data entry, then a storage routine will be needed to end the operation.

FUNCTION : XEQ A

(5) To run the first part of the regression program.

(5.1) Load the routine "R21".

INPUT : [ALPHA] R21 [ALPHA]

FUNCTION : XEQ READP

(5.2) Run the program.

FUNCTION : XEQ R21

(5.3) DISPLAY : FILENAME?

INPUT : data filename.

FUNCTION : R/S

DISPLAY : NDATA= (# of data points)

FUNCTION : R/S

DISPLAY : NVAR= (# of variables)

FUNCTION : R/S

DISPLAY : an audible beep.

Note: In the last step data files "TS" and "MTS" were created in the extended memory.

(5.4) To start the regression. If the user is carrying out another regression (with the same data) and has already given the information needed in steps (5.2) and (5.3) during the first pass, he may skip steps (5.2) and (5.3).

FUNCTION : XEQ A

DISPLAY : K?

INPUT : # of regression variables (excluding the dependent variable)

FUNCTION : R/S

(5.5) A loop will start to enable the user to select the "input" variables that will define the regression variables.

DISPLAY : VARi?

INPUT : j (such that regression variable i = input variable j)

FUNCTION : R/S

(5.6) The above loop will end by a request to define the dependent variable.

DISPLAY : Y?

INPUT : 1 (Y = input variable 1)

(5.7) A loop will begin to request the transformations that would be carried out on each regression variable (see table 1).

DISPLAY : TRNF Vi?

INPUT : (Optional) transformation code (If none make no input)

FUNCTION : R/S

(5.8) The above loop will end with a request to input the optional transformation for Y.

DISPLAY : TRNF Y?

INPUT : (Optional) transformation code for Y

FUNCTION : R/S

(5.9) The program will request the range of data to be entered in the regression. No input means to process all the data.

DISPLAY : NDATA?

INPUT : (Optional) bbb.eee control word to indicate that data point # bbb through eee will be processed.

(5.10) The model considered by the regression will be printed to remind the user of the selected variables and their transformations.

DISPLAY : depends on the model

(6) Program "R22" will be loaded and executed automatically at this stage to continue solving the regression. The output will be as follow :

```
(6.1) The coefficient of determination, R^2, is first shown.
DISPLAY : R2= (R^2)
FUNCTION : R/S*
(6.2) The regression coefficients are displayed in a loop.
DISPLAY : b1= (b1)
FUNCTION : R/S*
  ..........
FUNCTION : R/S*
DISPLAY : bk= (bk)
FUNCTION : R/S*
DISPLAY : b0= (b0)
FUNCTION : R/S*
(6.3) The ANOVA table is shown next.
DISPLAY :  ANOVA
FUNCTION : R/S*
DISPLAY : TOTAL CORR
FUNCTION : R/S*
DISPLAY : D.F.= (degree of freedom)
FUNCTION : R/S*
DISPLAY : SS= (sum of squares)
FUNCTION : R/S*
DISPLAY : REGRESS.
FUNCTION : R/S*
DISPLAY : D.F.= ( degree of freedom)
```

FUNCTION : R/S*

DISPLAY : SS= (sum of squares)

FUNCTION : R/S*

DISPLAY : RESIDUAL

FUNCTION : R/S*

DISPLAY : D.F.= (degree of freedom)

FUNCTION : R/S*

DISPLAY : SS= (sum of squares)

FUNCTION : R/S*

DISPLAY : F = (Snedocor-F statistic)

FUNCTION : R/S*

DISPLAY : S^2 = (regression variance)

FUNCTION : R/S*

(6.4) A loop will show the standard errors for the regression coefficients and their corresponding values of the critical student-t statistics.

DISPLAY : Sbi= (Sbi)

FUNCTION : R/S*

DISPLAY : Ti= (critical student-t)

FUNCTION : R/S*

(7) To obtain projection and confidence intervals use program "R23".

(7.1) Load program "R23".

INPUT : [ALPHA] R23 [ALPHA]

FUNCTION : XEQ READP

(7.2) To initialize the program.

FUNCTION : XEQ R23.

(7.3) to carry out projections.

FUNCTION : XEQ A

(7.3.1) The program will scan to the selected variables and prompt once for each variable even if it is involved in more than one term.

DISPLAY : Xi?

INPUT : Xi

FUNCTION : R/S

(7.3.2) The value of the projectd Y is shown.

DISPLAY : Y.= (Y hat)

(7.3.3) For another projection go to step (7.3).

(7.4) To carry out confidence interval calculations.

FUNCTION : XEQ B

DISPLAY : PROB?

INPUT : Confidence probability (fraction)

FUNCTION : R/S

DISPLAY : T= (student-t)

FUNCTION : R/S*

(7.5) A loop will start to show the upper and lower limit for the confidence intervals.

DISPLAY : bi

FUNCTION : R/S*

DISPLAY : UL= (upper limit)

FUNCTION : R/S*

DISPLAY : LL= (lower limit)

(8) For a new case or new model go to step (5).

(9) To input a new set of data go to step (4).

TABLE1

TRANSFORMATION CODES

tranformation	code
none	none
ln var	LN
1/var	1/
SQRT var	SQRT
var^2	SQR or P2
var^-2	1/SQR
var^-1/2	1/SQRT
var^3	P3
var^4	P4
var^5	P5
var^6	P6
var^7	P7

MASS STORAGE DATA CONFIGURATION

rec#	content
00	status filename
01	max # of data points
02	# of data points
03	# of variables
04	X1,1
05	X2,1
05	etc.

EXAMPLE

The following data are to be fitted using the following model:

$$X4 = b0 + b1\ X1 + b2\ X2 + b3\ X3$$

The data points are:

\ |	1	2	3	4	5
X1	7	1	11	11	7
X2	25	29	56	31	52
X3	6	15	8	8	6
X4	60	52	20	47	33

(1) Set the size to 100 to make sure that programs can be read. Use FIX 5.

A printer is recommended.

(2) Load "LIB2" and carry out a GTO.. to protect "LIB2" from being erased every time another program is read.

(3) Load "R2I" and run using the following information:

(3.1) Assign "DR2" as the data filename.

(3.2) Use default status filename (i.e. make no input for the status prompt).

(3.3) Let the maximum # of data points be 5, and the number of variables be 4.

(3.4) Key in the data with the aid of prompts.

(4) Load program "R21" and carry out the following:

(4.1) Run program and use the data file "DR2" we just created.

(4.2) Let k=3 and assign regression variables Var1, Var2 and Var3 to be X1, X2 and X3 respectively. Let Y be X4.

(4.3) Since the model tested is perfectly linear no transformations are needed for any term. Make no input when prompted for the transformations.

(4.4) When prompted for the data range make no input to signal that all the data should be involved.

(4.4) Wait for the results to come out.

(5) Load program "R23" and carry out the following.

(5.1) Initialize the program.

(5.2) Calculate the estimated Y (X4) when X1=1, X2=29 and X3=15. Use (LBL A).

(5.3) At 95 percent confidence calculate the student-t statistics and the confidence intervals for b1, b2 and b3.

```
                SIZE 100
                  FIX 5

LIB2
                  READP


                  GTO ..
PACKING

R2I
                  READP

                XEQ "R2I"
FILENAME?
DR2                 RUN
STATUS?
                    RUN
MAX DATA?
          5.00000   RUN
NVAR?
          4.00000   RUN
X1?
               7    RUN
X2?
              25    RUN
X3?
               6    RUN
X4?
              60    RUN
                    RUN
X1?
               1    RUN
X2?
              29    RUN
X3?
              15    RUN
X4?
              52    RUN
                    RUN
X1?
              11    RUN
X2?
              56    RUN
X3?
               8    RUN
X4?
              20    RUN
                    RUN
X1?
              11    RUN
X2?
              31    RUN
X3?
               8    RUN
X4?
              47    RUN
                    RUN
X1?
               7    RUN
X2?
              52    RUN
X3?
               6    RUN
X4?
              33    RUN
FILE FULL

R21
                  READP

                XEQ "R21"
FILENAME?
DR2                 RUN
NDATA= 5.00000
                    RUN
NVAR= 4.00000
                    RUN

                XEQ A
K?
          3.00000   RUN
VAR1?
               1    RUN
VAR2?
               2    RUN
VAR3?
               3    RUN
Y?
               4    RUN
TRNF Y1?
                    RUN
TRNF Y2?
                    RUN
TRNF Y3?
                    RUN
TRNF Y?
                    RUN
NDATA?
                    RUN
```

```
 X4 = b0
  + b1  X1
  + b2  X2
  + b3  X3

R2= 0.99894

b1= -1.28410
b2= -1.03693
b3= -1.33949
b0= 103.44732

   ANOVA
TOTAL CORR
D.F.= 4.00000
SS= 1013.20000

REGRESS.
D.F.= 3.00000
SS= 1012.12319

RESIDUAL
D.F.= 1.00000
SS= 1.07681

F = 313.30947
St2= 1.07681

Sb1= 0.18729
T1= -6.85636

Sb2= 0.03988
T2= -26.00001

Sb3= 0.19825
T3= -6.75658
```

```
R23
                    READP

               XEQ "R23"
                   XEQ A
X1?
                 1    RUN
X2?
                29    RUN
X3?
                15    RUN
Y. = 52.00000

                   XEQ B
PROB?
           .95000    RUN
T= 11.31131

b1
UL= 0.83435
LL= -3.40254
b2
UL= -0.58581
LL= -1.48804
b3
UL= 0.90297
LL= -3.58195
```