DOCUMENT RESUME

ED 090 284 TM 003 563

AUTHOR Mason, Robert L.; McNeil, Keith A.
TITLE MASHIT-For Ease in Regression Program Communication.
PUB DATE Apr 74
NOTE 18p.; Paper presented at the Annual Meeting of the American Educational Research Association (Chicago, Illinois, April 15-19, 1974)

EDRS PRICE MF-$0.75 HC-$1.50 PLUS POSTAGE
DESCRIPTORS Analysis of Covariance; Computational Linguistics; *Computer Programs; Correlation; Hypothesis Testing; Input Output; Multiple Regression Analysis; *Programing Languages; *Research; *Statistical Analysis
IDENTIFIERS FORTRAN; SNOBOL

ABSTRACT

This regression system is an intermediate result of a project to develop a comprehensive regression computer system as a foundation for a complete statistical man-machine interface. The outstanding features of the system can be condensed into two principal concepts. First, the program dynamically allocates core resulting in no limits on title cards, question cards, etc. Secondly, "English type" user commands are used in a free format mode to save computer instruction time. The resulting system has two phases, constructed in such a manner that additional capabilities can be added efficiently. A summary of the users manual is in the document. (Author/BB)

4.04

# MASHIT - FOR EASE IN REGRESSION

## PROGRAM COMMUNICATION

Robert L. Mason

Science Applications, Incorporated
2109 W. Clinton Avenue
Huntsville, Alabama 35805

Keith A. McNeil

Educational Monitoring Systems



Paper Presented to 1974 Annual Meeting Of
American Education Research Association
Chicago, Illinois

# ABSTRACT

## MASHIT - For Ease In Regression Program Communication

Robert L. Mason
Science Applications, Incorporated
2109 W. Clinton Avenue
Huntsville, Alabama 35805

This regression system is an intermediate result of a project to develop a comprehensive regression computer system as a foundation for a complete statistical man-machine interface. The outstanding features of the system can be condensed into two principle concepts. First, the program dynamically allocates core resulting in no limits on title cards, question cards, etc. Secondly, "English type" user commands are used in a free format mode to save computer instruction time. The resulting system is two phase constructed in such a manner that additional capabilities can be added efficiently.

MASHIT (Mason's Automatic Statistical Hypothesis Interpreter and Tester) is a computer program written in high level programming languages to facilitate the interaction between the computer and the researcher. The primary unique aspect of MASHIT is that the program performs regression analysis based on conversation language research hypotheses. Ultimately, other statistical techniques will be incorporated into the system; however, the current version handles that wide range of research hypotheses that can be tested using MLR. Indeed, any least squares hypotheses concerned with a single criterion can be tested with MASHIT. The program readily handles "analysis of covariance" questions.

After studying several available regression programs, a composite of shortcomings was compiled. There was no one system that offered all the features that the researchers and students desired to test their hypotheses. In addition, the only systems that allowed free form input were the interactive terminal programs, whereas the majority of researchers must cope with "batch mode" computer systems. MASHIT is a system directed toward researchers desiring ease and flexibility in accessing a "batch mode" computer system. The following is a list of the outstanding program features:

(a) Analysis of natural language regression questions.
(b) Free format (no column restrictions with the one exception of any optional FORTRAN transformation statements desired).
(c) Virtually unlimited number of variables.
(d) Virtually unlimited number of models.
(e) Virtually unlimited number of research questions.
(f) Virtually unlimited number of title cards.
(h) Any size variable labels.
(i) The program will dichotomize all "A" or "T" field (discrete) variables. The user does not have to keep track of newly created variables.
(j) Any FORTRAN transformation statements allowed.
(k) No parameter card necessary.
(l) All double precision calculations.
(m) Multiple returns from transformation subroutine.

MASHIT was written for the IBM 360/370 series computers. Constructed in two parts, the program first reads and analyzes the researchers instructions, and passes this information to the second stage which performs the regression and test calculations.

The first stage actually creates the second stage program resulting in a "tailored" regression program for each individual user. Storage sizes can be expanded or contracted to fit the researchers program requirement and the desired machine region (core) request. This flexibility of the program does not affect the number of variables which can be processed, however, the machine CPU time is decreased as the core storage size is increased.

In order for a computer program to interpret natural language, established criteria must be met. They are (1) variables must be prelabeled if referenced by labels in a hypotheses, (2) only specific phrases from a list of keyword can be used and (3) certain syntactical rules must be followed. These rules are discussed in a later section.

The flexibility of the program allows the user to input his "control deck" as though it were written in a manner similar to a paragraph. There are no column restrictions with the one exception of FORTRAN trnasformations. MASHIT searches for keywords and labels; therefore, blanks are placed between coded words.

The program reads the entire "control deck" as if it were one long card. Therefore, coding can skip from card to card, even with the option of inserting blank cards in the control deck. Slashes, periods, or question marks are delimiters indicating the end of one type of control information within the control deck. There are presently eight types of control cards as follows:

(a) Title Card(s)
(b) Label Card(s)
(c) Transformation Card(s)
(d) Special Command Card(s)
(e) Question Card(s)
(f) Model Card(s)
(g) Test Card(s)
(h) Format Card(s)

All the control cards are optional with the exception of a format. If the format is the only card included, MASHIT prints the means, standard deviations and the correlations. A summary of the rules for each type control card is included as a "mini reference guide" at the end of this paper. To facilitate the remainder of the discussion, example deck setups are shown to illustrate the features of MASHIT.

Example 1

This is a rather limited application of MASHIT. One question is asked without the use of labels. The program recognizes the word "PREDICT" and uses the variable that follows as the criterion. The first variable and the unit vector are employed as predictor variables. Since there are no covariates, the restricted model is inferred to have an $R^2$ value of zero. The number of subject and numbers of linearly independent vectors are calculated and the F test is evaluated. The next page contains the printout as generated by the program.

```
// (Job Card)
// EXEC  MASHIT
Does X1 PREDICT VAR 2?
(2F3.0)
```

Data Cards

```
/*
```

NUMBER OF OBSERVATIONS———————— 12

NUMBER OF VARIABLES READ———————— 2

NUMBER OF VARIABLES AFTER TRANSFORMATION—— 2

NUMBER OF VARIABLES - CONTINUOUS———— 2

NUMBER OF VARIABLES - DISCRETE———— 0

INPUT UNIT NUMBER———————— 5

FORMAT (2F3.0)

| VARIABLE NUMBER | TYPE OF VARIABLE | NUMBER OF DIFFERENT VALUES | MEAN | STANDARD DEVIATION | VARIABLE NAME |
|---|---|---|---|---|---|
| 1 | CONTINUOUS | | 31.50000 | 19.80951 | |
| 2 | CONTINUOUS | | 56.63333 | 26.73273 | |

CORRELATION MATRIX

| | 1 | 2 |
|---|---|---|
| VARIABLE 1 | 1.00000 | |
| VARIABLE 2 | -0.48373 | 1.00000 |

DOES X1 PREDICT VAR 2 ?

CRITERION NUMBER = 2 INDEPENDENT VECTORS = 2
MODEL R-SQUARE... = 0.23399740 NUMBER OF ITERATIONS = 1

| VARIABLE NUMBER | RAW SCORE WEIGHTS | VARIABLE NAME |
|---|---|---|
| 1 | -0.65279252 | |
| REGRESSION CONSTANT = | 77.39629787 | |

FULL MODEL..... = MODEL FROM ABOVE

RESTRICTED MODEL... = ZERO RSQ MODEL

$$F = \frac{(\text{RSQ F} - \text{RSQ R}) / \text{DF1}}{(100 - \text{RSQ F}) / \text{DF2}} = \frac{(0.23400 - 0.0) / 1}{(1.0 - 0.23400) / 10} = 3.054787$$

NONDIRECTIONAL PROBABILITY = 0.1084337

DIRECTIONAL PROBABILITY = 0.0542168
(IN HYPOTHESIZED DIRECTION)

Example 2

This example illustrates the use of the "title", "model", "label", and "test" cards. Note that the title is two cards in length and each variable label is placed on a seperate card. This is not necessary as a title can be any number of cards in length and labels only have to be seperated by a delimiter.

```
// (Job Card
// EXEC   MASHIT
THIS STUDY USES LABEL CARDS, TWO MODEL CARDS, AND
A TEST CARD /
LABELS:
  RUNNING SPEED FOR 100 YARD DASH:
  AGE:
  MOTIVATION /
MODEL  A:  AGE AND MOTIVATION PREDICTING RUNNING SPEED
           FOR 100 YARD DASH /
MODEL  B:  AGE PREDICTING RUNNING SPEED FOR 100 YARD
           DASH /
TEST MODEL A AGAINST MODEL B /
(3F3.0)
        Data Cards
/ *
```

The structure in this example is similar to that used in other hypothesis testing regression programs with the exception that the variables have been labeled and referenced in natural language in the delineation of the models. This technique is used when testing many restricted models against the same full model. A simpler structure is available especially if only one F test were being computed. By use of the "question" card as in example 1 with the attachment of a covariate phase, one question can replace two models and a test as the following:

DOES MOTIVATION PREDICT RUNNING
SPEED FOR 100 YARD DASH OVER AND ABOVE AGE?

Since all least squares hypothesis can be phrased in "covariance" terminology as above, the "question" card has much potential for researchers. The printed output is shown on the next two pages.

THIS STUDY USES LABEL CARDS, TWO MODEL CARDS, AND
A TEST CARD /

NUMBER OF OBSERVATIONS———————— 12
NUMBER OF VARIABLES READ———————— 3
NUMBER OF VARIABLES AFTER TRANSFORMATION— 3
NUMBER OF VARIABLES - CONTINUOUS———— 3
NUMBER OF VARIABLES - DISCRETE———— 0
INPUT UNIT NUMBER———————— 5

FORMAT (3F3.0)

| VARIABLE NUMBER | TYPE OF VARIABLE | NUMBER OF DIFFERENT VALUES | MEAN **** | STANDARD DEVIATION | VARIABLE NAME |
|---|---|---|---|---|---|
| 1 | CONTINUOUS | | 490.16667 | 245.29570 | RUNNING SPEED FOR 100 YARD DASH |
| 2 | CONTINUOUS | | 484.41667 | 265.58597 | AGE |
| 3 | CONTINUOUS | | 642.08333 | 274.91831 | MOTIVATION |

CORRELATION MATRIX
**********

| | 1 | 2 | 3 |
|---|---|---|---|
| VARIABLE 1 | 1.00000 | | |
| VARIABLE 2 | 0.26106 | 1.00000 | |
| VARIABLE 3 | 0.07457 | 0.21081 | 1.00000 |

MODEL A : AGE AND MOTIVATION PREDICTING RUNNING SPEED FOR 100 YARD DASH /

CRITERION VARIABLE NAME = ..... RUNNING SPEED FOR 100 YARD DASH

CRITERION NUMBER = 1 INDEPENDENT VECTORS = 3
MODEL R-SQUARE...= 0.06855198 NUMBER OF ITERATIONS = 2

| VARIABLE NUMBER | RAW SCORE WEIGHTS | VARIABLE NAME |
|---|---|---|
| 2 | 0.27580396 | AGE |
| 3 | 0.02121643 | MOTIVATION |
| REGRESSION CONSTANT = | 342.93861958 | |

MODEL B : AGE PREDICTING RUNNING SPEED FOR 100 YARD DASH /

CRITERION VARIABLE NAME = ..... RUNNING SPEED FOR 100 YARD DASH

CRITERION NUMBER = 1 INDEPENDENT VECTORS = 2
MODEL R-SQUARE...= 0.06815249 NUMBER OF ITERATIONS = 1

| VARIABLE NUMBER | RAW SCORE WEIGHTS | VARIABLE NAME |
|---|---|---|
| 2 | 0.28043419 | AGE |
| REGRESSION CONSTANT = | 354.31967218 | |

TEST MODEL A AGAINST MODEL B /

FULL MODEL.....= MODEL A

RESTRICTED MODEL...= MODEL B

$$F = \frac{(RSQ\ F - RSQ\ R)\ /\ DF1}{(1.0 - RSQ\ F)\ /\ DF2} = \frac{(0.06855 - 0.06815)\ /\ 1}{(1.0 - 0.06855)\ /\ 9} = 0.003860$$

NONDIRECTIONAL PROBABILITY = 0.9505624

DIRECTIONAL PROBABILITY = 0.4752812
(IN HYPOTHESIZED DIRECTION)

Example 3

Here is featured the input of nominal data, FORTRAN transformations and a "covariate" question. Three variables are read and a fourth is created as the square of the third variable.

```
// (Job Card)
// EXEC   MASHIT
        X (4) = X (3) ** 2
IS VAR 1 PREDICTED BY X 2
GIVEN KNOWLEDGE OF X (3), X 4 ?
(F5.0, A1, 4X, F5.0)
            Data Cards
/ *
```

Also, the A-format for variable two implies that it is discrete. MASHIT then automatically constructs and maintains the mutually exclusive group membership vectors. When variable two is referenced in the research question, the group membership vectors are substituted.

Notice in the printout (next pages) that the program reports the number of observations, how many variables were read, how many were created by transformations and the number of mutually exclusive vectors that resulted.

Further Documentation

MASHIT was developed by Robert L. Mason as part of a doctoral dissertation under the direction of Dr. Keith McNeil. The dissertation (Mason, 1973) has complete documentation. Also, a 65 page "MASHIT) user's guide is available. The following is a summary of the users manual.

NUMBER OF OBSERVATIONS———————— 15

NUMBER OF VARIABLES READ———————— 3

NUMBER OF VARIABLES AFTER TRANSFORMATION— 4

NUMBER OF VARIABLES - CONTINUOUS———————— 3

NUMBER OF VARIABLES - DISCRETE———————— 1

INPUT UNIT NUMBER———————— 5

NUMBER OF DICHOTOMIZED VARIABLES———————— 2

FORMAT (F5.0,A1,4X,F5,0)

| VARIABLE NUMBER | TYPE OF VARIABLE | NUMBER OF DIFFERENT VALUES | MEAN **** | STANDARD DEVIATION | VARIABLE NAME |
|---|---|---|---|---|---|
| 1 | CONTINUOUS | | 41.66667 | 29.12883 | |
| 2 | DISCRETE | 2 | | | |
| 3 | CONTINUOUS | | 53.20000 | 21.22326 | |
| 4 | CONTINUOUS | | 3280.66667 | 2415.94964 | |

| VARIABLE NUMBER | TYPE OF VARIABLE | CREATED BY VARIABLE NUMBER | MEAN **** | STANDARD DEVIATION | VARIABLE VALUE |
|---|---|---|---|---|---|
| 5 | DICHOTOMOUS | 2 | 0.66667 | 0.47140 | A |
| 6 | DICHOTOMOUS | 2 | 0.33333 | 0.47140 | B |

CORRELATION MATRIX

| | 1 | 2 | 3 | 4 | 5 | 6 |
|---|---|---|---|---|---|---|
| VARIABLE 1 | 1.00000 | | | | | |
| VARIABLE 2 | ******** | ******** | | | | |
| VARIABLE 3 | -0.38035 | ******** | 1.00000 | | | |
| VARIABLE 4 | -0.32130 | ******** | 0.96756 | 1.00000 | | |
| VARIABLE 5 | 0.03560 | ******** | -0.10662 | -0.17225 | 1.00000 | |
| VARIABLE 6 | -0.03560 | ******** | 0.10662 | 0.17225 | -1.00000 | 1.00000 |

**********

IS VAR 1 PREDICTED BY X 2 GIVEN KNOWLEDGE OF X ( 3 ) , X# ?

**********

CRITERION NUMBER = 1 INDEPENDENT VECTORS = 4
MODEL R-SQUARE...= 0.18111067 NUMBER OF ITERATIONS = 5
**********

| VARIABLE NUMBER | RAW SCORE WEIGHTS | VARIABLE NAME |
|---|---|---|
| 2 - 5 | 3.08248870 | |
| 2 - 6 | 0.0 | VALUE = A<br>VALUE = B |
| 3 | -1.55778456 | |
| 4 | 0.00946970 | |
| REGRESSION CONSTANT = | 91.41887092 | |

**********

CRITERION NUMBER = 1 INDEPENDENT VECTORS = 3
MODEL R-SQUARE...= 0.17883439 NUMBER OF ITERATIONS = 2
**********

| VARIABLE NUMBER | RAW SCORE WEIGHTS | VARIABLE NAME |
|---|---|---|
| 3 | -1.49366987 | |
| 4 | 0.00882175 | |
| REGRESSION CONSTANT = | 92.18867231 | |

**********

FULL MODEL.....= MODEL FROM ABOVE

RESTRICTED MODEL...= MODEL FROM ABOVE

$$F = \frac{( RSQ\ F - RSQ\ R ) / DF1}{( 1.0 - RSQ\ F ) / DF2} = \frac{( 0.18111 - 0.17883 ) / 1}{( 1.0 - 0.18111 ) / 11} = 0.030577$$
********

NONDIRECTIONAL PROBABILITY = 0.8583112

DIRECTIONAL PROBABILITY = 0.4291556
(IN HYPOTHESIZED DIRECTION)

**********

# "MASHIT" MINI GUIDE

Program MASHIT (Mason's Automatic Statistical Hypothesis Interpreter and Tester) was developed to aid the researcher in his ever losing battle with the computer. The program is written in SNOBOL and FORTRAN IV to run on the IBM 360-370 series computers. Presently only regression type questions can be interpreted.

This Mini reference guide is intended to be a summary for the computer user. If questions arise, the user should refer to the "MASHIT" users manual, which is complete with examples (Mason, 1973).

## CONTROL CARD RULES

The program looks for keywords in the statement made to the computer. Care with the spelling and spacing of input words is necessary. Space must be maintained between words and data input can be continued from card to card freely as the computer thinks the deck is one long card. The following are general types of cards with the limited rules necessary. An important point to note is that the only card absolutely necessary besides the data is a FORMAT card. Just placing the FORMAT card before the data results in the printing of means, standard deviations, and correlations.

A. TITLE CARD(S) - Optional
   1. Any number of cards
   2. Must use keyword "TITLE" or "PROJECT" or "STUDY".
   3. Must end with slash or period.

B. TRANSFORMATIONS - Optional
   1. The rules of FORTRAN apply.
   2. Refer to variables in the Array X.

      EXAMPLE: X(22) = X(1) *X(3)

C. LABEL CARD(S) - Optional

1. Must start with keyword "LABEL" or "LABELS" followed by a colon or semicolon.
2. Separate names by colons or semicolons.
3. Can indicate or change variable number. Otherwise they are thought to be sequential. This example labels variables 1, 2, 10 and 11.

   LABEL: SEX: RACE:
   VAR 10: GRADE POINT: EDUCATION /

4. Must end entire variable labeling series with a slash or period.

D. SPECIAL INSTRUCTIONS - Optional

1. The following are instructions that the program understands:
   a. READ IN 20 VARIABLES
   b. READ DATA FROM UNIT 4
   c. READ IN 120 OBSERVATIONS
   d. REGION SIZE = 132K
   e. THERE ARE 514 VARIABLES AFTER TRANSFORMATIONS.
2. Must end with slash or period.

E. MODEL(S) - Optional

1. Must have a model name that includes keyword "MODEL".
2. Model name must be followed by colon or semicolon.
3. Model structure follows the colon delimiter. Model structure is discussed later.
4. Model structure must end with a period or slash.

F. TEST CARD(S) - Optional

1. Must have keywords "TEST" and "AGAINST" or "WITH".
2. Must have at least two model names that were previously defined.
3. Must end with period or slash.

G. QUESTION(S) - Optional

1. Must use the model structure that is defined elsewhere.
2. Must end with period or slash.

H. FORMAT CARD(S) - Necessary

1. Can start the card with "(" or "FORMAT (".
2. Must have balanced parentheses.
3. Program will count the number of variables by the format.
4. F type variables considered continuous variables.
5. A type or I type variables are considered discrete and are dichotomized for the researcher.

## MODEL STRUCTURE

All questions and models must be formed in regression terms. That is, all predictor (independent) variables and the criterion (dependent) variables must be separated by a key word or phrase. Examples of this are:

A. DOES SEX, EDUCATION LEVEL, AND IQ PREDICT GRADE POINT AVERAGE?

B. VAR 5, VAR 7 AND VAR 9 PREDICTING X1.

C. IS THE VARIANCE IN X1 ACCOUNTED FOR BY SEX?

The keywords are underlined in the examples. The present list of keyphrases consists of:

| | |
|---|---|
| PREDICT | EXPLAIN |
| PREDICTS | EXPLAINS |
| PREDICTING | EXPLAINING |
| PREDICTIVE OF | EXPLAINED BY |
| PREDICTION OF | ACCOUNT FOR |
| PREDICTED BY | ACCOUNTED FOR BY |
| INFLUENCED BY | |

## OTHER CONCEPTS

Two other concepts regarding the structure must be mentioned. The first is that of covariates, and the second is naming and grouping variables. In asking a question, input can include covariate variable(s) in the question by the use of one of the following keyphrases:

OVER AND ABOVE
WITH KNOWLEDGE OF
GIVEN KNOWLEDGE OF
IN CONJUNCTION WITH
IN THE PRESENCE OF
WITH - list - AS COVARIATES

An example is:

DOES GROUP MEMBERSHIP PREDICT VAR 1 OVER AND ABOVE SEX?

The other concept is the naming and grouping of variables. Variables can be denoted by an assigned name, or using "X" or "VAR" notations. A comma or the word "AND" between two variables names means to use only those two variables. A dash (-) or one of the keywords (TO, THRU, THROUGH) indicates the use of those variables and all variables between. The example:

VAR 7, VAR 9 - 11

refers to the four variables 7, 9, 10, 11. For further information and examples, the user is referred to "MASHIT" users manual (Mason, 1973) describing the program.

## DECK SETUP

The following card order is used in making a run on program MASHIT.

```
// (Job Card)
// EXEC   MASHIT
                    Consists of the following types of cards
                    in any order:
                    A.   TITLE
                    B.   TRANSFORMATIONS
CONTROL DECK        C.   LABELS
(All Optional)      D.   INSTRUCTIONS
                    E.   MODELS
                    F.   TESTS
                    G.   QUESTIONS
FORMAT CARD(S)

        Data Cards

/ *
```

## Example

```
// (JOB CARD)
// EXEC   MASHIT
THIS STUDY MEASURES THE CURVILINEAR EFFECT OF AGE
ON RUNNING SPEED /
LABELS :
  RUNNING SPEED FOR 100 YARD DASH :
  AGE :
  AGE SQUARED, USED FOR CURVILINEAR TEST /
     X(3) - X(2) ** 2 /
MODEL A : VAR 2, VAR 3 PREDICTING VAR 1 /
MODEL B : VAR 2        PREDICTING VAR 1 /
TEST MODEL A AGAINST MODEL B /
DOES X3 PREDICT VAR 1 OVER AND ABOVE AGE?
(2F3.0)

        Data Cards

/ *
```

NOTE: The one question does the same thing as the two models and test combined.

## REFERENCES


Mason, R. L. "The Development of an Automated Statistical Hypothesis Interpretor and Tester". Unpublished doctor's dissertation, Southern Illinois University Carbondale, Illinois, 1973.

Mason, R. L. "MASHIT" User's Manual, unpublished manuscript, 1973.