Errata

The *WATFIV NOLIST and *DATA cards should be omitted from your deck, as they have been already included in the card files.

NOTE: MAI2 is currently not available.

Use of the IT Regression Package TOAD

P. Firey

BU-654-M
September, 1978

TOAD is a small computer program which can do multiple regressions, plots, summary statistics, PRESS variable selection, and Cp variable selection. Its primary use is as a teaching aid. It is designed to run on Cornell's free system, IT, and was written by Chuck Bremer in the computer language FORTRAN.

This explanation of the use of TOAD assumes only two pieces of knowledge, namely: one, how to keypunch; and two, how to submit an IT job to the Cornell computer. If either of these procedures is new to you, please consult your TA, the instructor, or someone at the computer terminal.

First, here is a brief word on precisely what TOAD can do and some comparison with more advanced programs.

TOAD can read in up to fifteen variables, and can do multiple regression on up to eleven variables, including the dependent variable and the intercept, and it will easily handle up to fifty observations. PRESS and Cp selection may be done on data sets where the number of variables times the number of observations is less than 400. It can plot any variable against any other, and it is possible to do residual and predicted plots. An option has been included in the program to allow data sets with more observations to be used, but limitations on how much can be done in a single job are set and no plots may be done. It can take as input the X'X matrix instead of the raw data, albeit again with some limitations. It will print summary statistics of the data, and if desired standardize (i.e., subtract the mean and divide by the standard deviation) the data before doing the regressions. TOAD can also do two types of variable selection, namely PRESS and Cp stepwise selection. TOAD also has the capacity to transform variables, and create new variables.

At this point, you may start to think that there must be a catch somewhere, since this seems like a lot to get for free. This is true; there is a catch, and that involves the relatively primitive character of the program.

To be more specific: for one thing, TOAD does not let you use much real English in its instructions; most directions to the program are numbers, with translations to English being given in the manual. For example, if you want to print out the data, you cannot say PRINT DATA, or any other English phrase. There is a way to print the data, but you signify it by punching a one instead of a zero on one of the cards. For another, in referring to the variables, numbers are used instead of names: i.e., the first variable on your data cards is called 1, the second 2, etc. These are the only labels you will see on the

output (that is, the printed material produced by the computer). You will have to remember what number corresponds to what variable. Thirdly, you will have to give the program quite a bit of information--in particular, how many variables there are, how many observations, and so forth. More sophisticated programs count them for you. Finally, while there are a few moderately informative messages that may appear if you have typed something wrong, by and large the consequence of a mistake will be the appearance of a mysterious message on your output that you will usually have trouble interpreting. If this happens, don't panic; check your cards. If you cannot find a keypunching or other error on them, check back with the TA, bringing both the cards and the output.

CARD DECK CONSTRUCTION

The first card of the deck should be a blue IT job card. This is followed by a card to tell the computer what language the program is in, which for our computer looks like the following:

*WATFIV NOLIST

The star must be in the first column of the card. The NOLIST is a message to the computer that you do not want a copy of the program printed out; this is necessary, since the program is quite long and if it is printed out there is no room for the printing of the statistical information which it calculates. Next follow several cards to fetch the program. Unfortunately, here we encounter another complication because we are getting this for free: not all the program will fit at one time into one job. Luckily, we can get several useful combinations of jobs. For example, we can do regression and plotting together, plotting and summary statistics together, regression and standardization together, or just variable selection. The following cards are the ones necessary to do regression and plotting.

/*INSERT PJQ.MAIN
/*INSERT PJQ.REGR
/*INSERT PJQ.PLOT
/*INSERT PJQ.ENDD

The following cards are the ones for plotting and summary statistics.

/*INSERT PJQ.MAIN
/*INSERT PJQ.STAT
/*INSERT PJQ.PLOT
/*INSERT PJQ.ENDD

The following cards are the ones for standardization and regression.

/*INSERT PJQ.MAI2
/*INSERT PJQ.NORM
/*INSERT PJQ.ENDD

Finally, these are the cards for variable selection.

/*INSERT PJQ.MAIN
/*INSERT PJQ.PRES
/*INSERT PJQ.ENDD

Please note each of these start with the same card, which fetches the part of the program which reads in the data, and ends with the same card, which gets the part of the program which finishes up the program, as well as other things.

EACH OF THESE CARDS MUST BE PUNCHED EXACTLY AS SHOWN, WITH THE / IN THE FIRST COLUMN! Next, after your choice of /*cards, henceforth called program cards, the following card must appear.

*DATA

Again, this is punched starting in the first column; this card is to tell the computer that data and instructions to TOAD are next.

TOAD CARDS

There are three basic types of cards read by TOAD: TASK cards, FORMAT cards, and VARIABLE LIST cards.

TOAD processes these cards in the order in which they appear in the deck; that is to say, data input and statistical procedures are performed in the order in which you request them. Note in particular that a procedure cannot make use of information which is produced by another procedure that is requested later; TOAD lacks the intelligence needed to look ahead in the deck. (There is a somewhat confusing exception to this rule, which arises in the processing of large data sets; the matter is discussed later in this write-up.)

IMPORTANT NOTE.
At the end of your card deck, after all the other cards, there must be a card to signify end of job to the computer. This is a // card, like this:

//

These //'s must appear in the first two columns of the card.

TASK cards are used to request the various procedures. They have the following form: A four letter mnemonic in the first four spaces of the card, six 'control' positions, I,II,through VI, ending in columns 14, 18, 22, 26, 30, and 34, and a title space, in columns 41 through 80. The possible four letter mnemonics are:

REGR	regression
DATA	data input
PRES	variable selection
STAT	summary statistics
NORM	standardization of data
PLOT	plotting
X-PR	X'X matrix input

The control positions are for numbers which relay information to the program. These must be punched so that the last number is in the last column of the position--i.e., if a 4 is to be punched, it should be in column 14, or column 18, etc. A 20 to be punched in position I should be punched with the 2 in column 13 and a 0 in column fourteen. What should be punched on each type of TASK card is listed below, under the individual descriptions. The title part of the card, from 41 through 80, is optional-- if you wish a label or title to appear connected with that section of printing associated with a particular task card, punch that title in the last 40 columns of that TASK card.If you leave it blank, no title will appear.

FORMAT cards are used to tell the program where each of your variables is punched on the data cards. (For the knowledgable, this is a standard FORTRAN format, enclosed in parentheses.) In most cases, this will be supplied to you for each data set by your TA. If you wish to use your own data, please consult

the TA for the proper form for this card. Usually, unless more than one set of data is used in one job, only one FORMAT card will appear in your deck.

VARIABLE LIST cards are cards containing a list of variables, and are required after some of the TASK cards. They allow you more flexibility, in that you may read in more variables in the data set than you wish to use in any of the procedures. They are lists of the variable numbers, that is, the ordering number of the variable according to the order it was on the data cards, and are punched so that the last digit of each number is in a column divisible by four--i.e., if you wished to specify the list 9, 2, 3, 4, and 12, they would be punched in columns 4, 8, 12, 16, and 19-20, respectively.

SPECIFIC PROCEDURES AND TASK CARDS

DATA

A DATA TASK card is used to read in a raw data set so that the program can then use it for calculations, plotting, etc. It should appear then, as the first TASK card, right after the *DATA card, unless one is directly reading in the X'X matrix. It has DATA punched in the first four columns, with the first and second positions having the following information.

POSITION I: the total number of observations.

POSITION II: the number of variables to be read for each observation.

Please note that usually each data point is on one card, with a data card for each observation.

IMPORTANT FACT: The program also generates a column of ones, an intercept column, for you. This need not be punched in the data set. The number of the intercept column is one plus the number of variables you have read in--i.e., if you read in four variables, an intercept column is available to you as variable 5.

After the DATA TASK card, a FORMAT card must appear with the format of your data. After this, put your data cards.

So, to read in a raw data set with three variables and 17 observations, with a format (given) oF F3.0,F4.0,F5.1, the following 19 cards would appear after the *DATA card.

```
DATA         17   3                    FIREFLY DATA-PG 10-3
(F3.0,F4.0,F5.1)
 45   26 21.1
 40   35 23.9
 58   40 17.8
 50   41 22.0
 31   45 22.3
 52   55 23.3
 54   55 20.5
 38   56 25.5
 40   70 21.7
 28   75 26.7
 38   79 25.0
 36   87 24.4
 36  100 22.3
 46  100 25.5
 40  110 26.7
 31  130 25.5
 40  140 26.7
```

REGR

REGR TASK cards actually request a multiple regression. There are a great many options concerning what portions of the calculation you wish printed out, and so the control positions may look very confusing at first. If you forget, and fail to type in numbers for all but the first (which is the only vital one), the program will still work, but you will only get a minimum of output-- to wit, an ANOVA table, and estimates for the betas. (This is true because of a technicality--the computer will interpret blanks as zeroes if it is expecting a number in that spot.)

First, in the first four columns there should be the letters REGR. Next, there should appear in position I the total number of variables in the calculation--i.e., the number of the indepedent variables, plus one (for the dependent variable) and plus another one if an intercept is going to be used. (If you were doing the Doolittle calculation by hand, this would be the number of columns you would write down for the calculations.) The title spot may be used to title the regression, and the other positions control the following:

Position I: (vital) the number of variables, as discussed above.

Position II: punch a 1 here if residuals and predicted values are desired. (If you want to do residual plots, a one must be punched here.) Otherwise, leave it blank or punch a zero.

Position III: this position indicates whether you wish to request any point estimates and standard errors for data points not in the data set, i.e., linear combinations of the b's. If you do not, leave this blank or type in a zero. If you do, punch in the number of point estimates you are requesting-1,2,3,etc. The coefficients or data for these points must be in the same format as the data, and are placed two cards after the REGR TASK card.

Position IV: A one punched here will print out the data used in the regression. Otherwise, leave it blank or punch a zero.

Position V: This is the position which controls how much of the calculation is printed out. If a 1 is punched here, the X'X matrix will be printed, and if a 2 is punched here, both the X'X matrix and the Doolittle calculation will be printed. If neither is desired, leave this blank or punch a zero.

Position VI: A one here will cause the inverse of the X'X matrix to be calculated and printed out. Do not do this unless you really need to, as it involves considerably more work for the computer and may limit the number of things you can do in one job.

After each REGR card must come a VARIABLE LIST card, which contains a list of the variables you want in the regression. If an intercept is desired, the number of the intercept variable must be first, and the number of the dependent variable must be last. Between these would be the independent variables, in the order in which you want them to appear in the model.

If you have indicated that you wish to do point estimates (something other that a zero or a blank in the third position) then the data cards with the coefficients must be put here, after the VARIABLE LIST card.

Examples are shown below for sets of cards to do regression on the firefly

data set. (Note here that variable 1 is the Y, variable 2 is X1, variable 3 is X2, and variable 4 is the intercept, or X0.)

If you wanted practically everything printed, the cards would be

```
REGR          4   1   0   2   1   0      REGR USING X0,X1,X2
   4   2   3   1
```

If you wanted a point estimate for the point X1=29, X2=22.0 the cards would be

```
REGR          4   1   |   2   1   0      REGR WITH POINT EST.
   4   2   3   1
 00  29 22.0
```

If you wanted just the ANOVA and b's, the following would suffice.

```
REGR          4                          SHORT REGRESSION
   4   2   3   1
```

And finally, if you didn't want the intercept in the model the cards would be as follows:

```
REGR          3                          SHORT REGR USING X1 X2
   2   3   1
```

You may run several REGR TASK cards (together with their VARIABLE LIST cards and coefficient cards, if needed) in one job.

PLOT

PLOT TASK cards are very simple. The first four columns contain the letters PLOT, position I contains the number of the variable to be plotted on the x axis, and position II contains the number of the variable to be plotted on the y axis. A title for the plot may be punched in columns 41-80, as before.

So, a PLOT TASK card to plot the Y of the firefly data against X2 would look like this:

```
PLOT          3   1                      PLOT OF Y VRS X2
```

PLOT TASK cards may go anywhere in the program after the DATA TASK card; thus, they may come either before or after REGR and STAT TASK cards. Of course, they must appear after the cards which produce whatever data is to be plotted; this point is discussed in more detail below when we consider residual plots.

SPECIAL INSTRUCTIONS FOR PLOTTING RESIDUALS AND PREDICTED VALUES.
If you have put a 1 in position II of a REGR TASK card and so requested that the residuals be calculated, the residuals and predicted values are then available

as variables 17 and 16, respectively. The values so calculated are ephemeral; that is, they are not punched on your data cards, and therefore may be used for plotting only in that same job. If a second REGR TASK card included in the same job also requests residuals, the same variable numbers are reused. To summarize, then, if you wish to do a residual plot, you use cards which look like the following:

```
REGR          4   1                    REGRESSION-X0,X1,X2
   4   2   3   1
PLOT         16  17                     RES. VRS PRED.-X0,X1,X2
```

Note that the PLOT TASK card here must come after the REGR TASK card, or else the PLOT card would be trying to use something that wasn't even computed yet.

Finally, to request residual plots for two regressions in the same job the cards would look like the following:

```
REGR          4   1                    REGRESSION OF FIREFLY DATA-X0,X1,X2
   4   2   3   1
PLOT         16  17                    PLOT OF RESIDUALS VRS PREDICTED-X0,X1,X2
REGR          3   1                    REGRESSION OF FIREFLY DATA-X0,X2
   4   3   1
PLOT         16  17                    PLOT OF RESIDUALS VRS PREDICTED-X0,X2
```

Note here that the PLOT TASK cards are after each REGR set.

STAT and NORM

Both STAT and NORM TASK cards print out summary statistics for selected variables; they differ in whether or not those variables are then standardized, i.e., the mean subtracted and the data divided by the standard deviation. The standardization of the data will persist for any further statistical procedures requested in that job. Whether you wish to do this or not is a statistical question, not a computational one; if in doubt, consult a statistician. In either case, the form of the card is the same, namely the four letters STAT or NORM in the first four columns, and the number of variables(up to a maximum of eight) you wish to select in position I. A title in columns 41 through 80 is allowed, and the STAT or NORM TASK card must be followed by a VARIABLES LIST card, containing the variable numbers of the variables selected. If standardization is desired as a preliminary to regression, be sure to use the /*INSERT PJQ.NORM card, which replaces both the /*INSERT PJQ.STAT and the /*INSERT PJQ.REGR card. If this is requested, the maximum number of variables used in one regression is reduced to 10, including the intercept and he dependent variable.

As an example of this, the cards to request summary statistics for the Y and X1 of the firefly data set would be punched as follows:

```
STAT          2                       SUMM. STAT. FOR Y AND X1
   1   2
```

The cards for just X2 would look like

```
STAT          1                       SUMM. STAT. FOR X2
```

3

PRES

Both Cp and PRESS variable selection are requested together, using a PRES TASK card; there is no way to request just one of them. The PRES TASK card has PRES in the first four columns, and has two control positions, as listed below. A title again is permissable in columns 41 through 80.

Position I: This position should contain the total number of variables in the selection, including the Y or dependent variable, and the intercept, if used.

Position II: If any variables are to be forced into the model, the number of forced variables should appear here. If none are to be forced, leave this column blank or punch a zero.

The PRES TASK card must be followed by a VARIABLES LIST card, listing the variable numbers to be used in the selection, with with the variable number of the dependent variable variable last. If you are forcing variables (position II non-negative) then this must ve followed by a second VARIABLE LIST card, listing the variables to be forced. To do PRESS and Cp on the firefly data set, then, (this, of course, is purely an academic exercise, as the data set contains so few variables) the following cards might be used, if one wished to include all the variables and force none:

```
PRES          4   0                        PRESS-NO VAR FORCED
   4   2   3   1
```

To do the same, but forcing in variable 2, X1, the cards become

```
PRES          4   1                        PRESS-ONE VAR FORCED
   4   2   3   1
   2
```

X-PR

the X-PR TASK card is used when you wish to do multiple regression with the X'X matrix as input. It takes the place of both the DATA and the REGR TASK cards, and combines features of both. It has X-PR in the first four columns, and the number of variables in position I. The number of observations that went into the computing of the X'X matrix is punched in position II, and positions III, V, and VI are as on the REGR TASK card. Position IV is ignored, as there is no data to print. A title, as usual, may be placed in columns 41 through 80. The X-PR TASK card must be followed by a FORMAT card, with the format of the X'X matrix. After this comes the X'X matrix, one row to a card. The matrix must be complete; that is, if an intercept is desired, it must be in the X'X matrix already. The y's must also be included, as the last column and row. If point estimates are requested (something other that a blank or zero in position III) these go directly after the X'X matrix, punched in the same format. Please note that no variable selection is possible; all the variables read in from the X'X matrix will be used in the regression, in the order listed.

An example of this, using the firefly data again, is as follows:

```
X-PR          4  17   1        2   0       X-PR READ
```

```
(4F10.2)
    17.00    1244.00     400.90     703.00
  1244.00  109328.00   30228.90   49557.00
   400.90   30228.90    9555.65   16363.10
   703.00   49557.00   16363.10   30211.00
```

Long data sets.
Data sets of more that fifty variables may be processed by TOAD, if certain modifications of the order in which the TASK cards are placed are made. To be more explicit: plotting is not allowed, and normalization of the data is not permitted. Furthermore, only one procedure, REGR, STAT, or PRES, is allowed in each job.

REGR: long data sets.
To do multiple regression on a long data set, the order of the cards changes to the following: (requests for residuals or for printing of the data will be ignored)

DATA TASK card
FORMAT card
REGR TASK card
VARIABLE LIST card
  data cards
ANY coefficient cards.

STAT: long data sets. To request summary statistics for the data, the order of the cards is changed to the following:

DATA TASK card
FORMAT card
STAT TASK card
VARIABLE LIST card
  data cards

PRES: long data sets.
PRES can take data sets up to a size limited by a maximum value of the product of the number of variables and the number of observations; this product must be less than 400. The order of the cards is

DATA TASK card
FORMAT card
PRES TASK card
VARIABLE LIST card
VARIABLE LIST card(if variables are to be forced)
  data cards

TRANSFORMATION OF VARIABLES.
This is perhaps the hardest part of TOAD to use, as it requires the user to actually write some FORTRAN lines. Be of stout heart; it is not as hard as it seems.

This is the main purpose of the part of the program called ENDD. It is a part called a subroutine which is called, or reread, every time TOAD reads in one observation. If a modification has been included, TOAD will do whatever data TRANSFORMATION the user desires for each observation. Some simple examples are given, and your instructor or a consultant can help you with others.

First, the program calls the data, reasonably enough, SET. It is stored as a matrix, with the variable number as the ith index and the observation number as the jth index, to use classical matrix algebra terminology. Unfortunately, often programmers are backward and refer to the second subscript as an i and the first as a j. This is true in this case. Therefore, if we wish to talk about the fourth variable, the program would like to call it
SET(4,I)
Similarly, the first variable is SET(1,I), the seventh SET(7,I), etc. By placing an I in the second spot, we are not specifying which observation, and so are referring to the entire variable.

Now, a quick word about computer functions--think of them as buttons on a calculator, and you won't be far off: the signs for addition, subtraction, and division are +, -, and / respectively. The less familiar ones: that for multiplication is *, and that to raise to a power is **. These signs are used to specify which transformations are desired. (Other functions are available, including log, exponential, sine, cosine, etc.; see your instructor if you are interested in these)

Now we are ready. To include a modification in ENDD, we do the following.
/*I PJQ.ENDD
is replaced by two cards,
/*I PJQ.ENDD,1-10
and
/*I PJQ.ENDD,11-END
with our modifications between them.

The modifications take the form
SET(new variable number,I)=some function of SET(old variable number,I)
The statement must start in the seventh column. The new variable number is the variable number where you wish the computed variable to be put, and should be so chosen as not to conflict with any of the variables you are reading in as input. If you are reading in four variables, remember that the intercept is created for you in 5, so you could start your created or transformed variables with 6. More than one old variable can appear on the right side of the equation; for example, a new variable may be the product of two old ones. As many statements of the above form may be included as you wish, as long as the total number of read-in variables and computed variables is less than or equal to fifteen.

Some common examples are as follows: (here we again use the firefly data set; with the intercept, four variables are already in use.)
Squares of variables.
In order to include X1 squared in the data set, the following is the correct

statement.

```
SET(5,I)=SET(2,I)**2
```

Cross products. In order to include X1*X2, the statement would be (assuming X1 squared is already using variable 5's spot)

```
SET(6,I)=SET(2,I)*SET(3,I)
```

These are only a few of the many things that may be done; essentially, any FORTRAN statements may be included as desired.

NOTE on order of operations done by the program: ALL data transformations are done line by line as the data is being read in, so all the variables created are available thereafter for multiple regression, statistics, variable selection, plotting, etc.

There is a summary of this handout accompanying the program, convering adaptation of the program to different systems. It is usually filed as WART. However, it is very concise, to the point of unreadability, and so is recommended only as a last resort.

Finally, as a summation, here is a sample job. This example does three plots, and two regressions.

```
IT JOB CARD
/*INSERT PJQ.MAIN
/*INSERT PJQ.REGR
/*INSERT PJQ.PLOT
/*INSERT PJQ.ENDD
DATA         17   3                    FIREFLY DATA-PG 10-3
(F3.0,F4.0,F5.1)
 45   26 21.1
 40   35 23.9
 58   40 17.8
 50   41 22.0
 31   45 22.3
 52   55 23.3
 54   55 20.5
 38   56 25.5
 40   70 21.7
 28   75 26.7
 38   79 25.0
 36   87 24.4
 36  100 22.3
 46  100 25.5
 40  110 26.7
 31  130 25.5
 40  140 26.7
PLOT          1   3                    PLOT OF Y AND X2
REGR          4   1                    REGRESSION OF FIREFLY DATA-X0,X1,X2
   4   2   3   1
PLOT         16  17                    PLOT OF RESIDUALS VRS PREDICTED-X0,X1,X2
REGR          3   1                    REGRESSION OF FIREFLY DATA-X0,X2
   4   3   1
PLOT         16  17                    PLOT OF RESIDUALS VRS PREDICTED-X0,X2
//
```