SAND79-1472
Unlimited Release

Stepwise Regression With PRESS and Rank Regression (Program User's Guide)

Ronald L. Iman, James M. Davenport, Elizabeth L. Frost, Michael J. Shortencarier

Prepared by Sandia Laboratories, Albuquerque, New Mexico 87185 and Livermore, California 94550 for the United States Department of Energy under Contract DE-AC04-76DP00789

Printed January 1980

When printing a copy of any digitized SAND Report, you are required to update the markings to current standards.

Sandia Laboratories

2900 Q(7-73)

Issued by Sandia Laboratories, operated for the United States Department of Energy by Sandia Corporation.

NOTICE

This report was prepared as an account of work sponsored by the United States Government. Neither the United States nor the Department of Energy, nor any of their employees, nor any of their contractors, subcontractors, or their employees, makes any warranty, express or implied, or assumes any legal liability or responsibility for the accuracy, completeness or usefulness of any information, apparatus, product or process disclosed, or represents that its use would not infringe privately owned rights.

SAND79-1472
Unlimited Release
Printed January 1980

Stepwise Regression with PRESS and Rank Regression
(Program User's Guide)

Ronald L. Iman
Statistics and Computing Division 1223
Sandia Laboratories
Albuquerque, New Mexico 87185

James M. Davenport
Department of Mathematics
Texas Tech University
Lubbock, Texas 79409

Elizabeth L. Frost
Statistics and Computing Division 1223
Sandia Laboratories
Albuquerque, New Mexico 87185

Michael J. Shortencarier
Statistics and Computing Division 1223
Sandia Laboratories
Albuquerque, New Mexico 87185

ABSTRACT

This report contains a description of a stepwise multiple regression program. This program provides for either a forward stepwise or backward elimination solution to multiple regression problems. The program also provides PRESS values that can be used for subset selection as well as providing options for regression analysis on the ranks of the data and for a weighted regression analysis on either raw or rank transformed data. This document has been written and designed for users of this STEPWISE program.

THIS PAGE LEFT BLANK INTENTIONALLY.

Table of Contents

		Page
I.	Introduction	1
II.	Output	2
III.	Input - Parameter Cards	2
	A. TITLE	2
	B. DATA	2
	C. LABEL	3
	D. BACKWARD	4
	E. FORCE	5
	F. OUTPUT	5
	G. PLOT RESIDUALS	6
	H. STEPWISE	6
	I. MODEL	7
	J. PRESS	9
	K. RANK REGRESSION	10
	L. WEIGHT	12
	M. DROP	12
	N. TRANSFORMATION	13
	O. END OF PARAMETERS	15
	P. FORMAT	15
	Q. END OF DATA	16
IV.	Additional Data Transformations	17
V.	Output of the Regression Coefficients and the Observations with their Predicted Values When Using Rank Regression	19
VI.	Deck Setup for STEPWISE	22
VII.	Deck Setup for Standardizing Data	24
VIII.	Bibliography	27
IX.	Example	28

values from rank regression analysis can also be saved on disk for additional investigation outside the STEPWISE program, (10) weighted regression analysis on the raw data or rank transformed data.

The reader who is familiar with SAND76-0364 (Iman, 1976) and has used STEPWISE on previous occasions is advised to read this report to note the extensive additions.

II. OUTPUT

The program will output means, variances, standard deviations, standard errors and coefficients of variation for each variable read in or generated via transformations. As options, the user may request the table of correlation coefficients, sum of squares and cross-products matrix, inverse correlation matrix and residuals y-values and $\hat{y}$'s, PRESS values and a plot of these PRESS values. The program will also print an analysis of variance table for a regression model and a table of statistics regarding the regression coefficients. Residual plots may be requested, and if they are, will be printed on the line printer.

III. INPUT - PARAMETER CARDS

All parameter cards must start in column 1 and may be placed in any order. An explanation and illustration of each of the parameter cards follows.

A. TITLE card (optional).

The "TITLE" card may contain any alphameric data that may be meaningful to the user. The word "TITLE" identifies the card, and its contents are printed at the top of each page of output. Only one "TITLE" card may be used.

Example:

```
TITLE, THIS IS A SAMPLE TITLE CARD
```

B. DATA card (required).

The "DATA" card has 3 arguments specified as follows:

```
DATA, NV, NT, DATDIS.
```

where

NV is the number of user supplied variables to be read into the programming;

NT is the number of additional variables created as transformations of other variables.

DATDIS is the data disposition parameter with the following codes:

= 0 the data are to be read from cards and not saved for subsequent use (this is the usual case and if not specified DATDIS will default to 0).

= 1 the data are to be read from cards and saved on disk (file 10 as binary records) for subsequent analysis of the data. (DATDIS would be set equal to 2 for any subsequent analyses following DATDIS = 1.)

= 2 the data are to be read from disk (file 10 - binary records).

In most cases the user need specify only the first argument (NV) and the other two will default to zero. The example below shows how this is done:

```
DATA, 10.
```

The above card indicates that 10 variables should be input from cards and not saved on disk. Both NT and DATDIS default to zero because they are not specified on the card. Note that the "DATA" card must be terminated by a period.

C. LABEL card (optional).

The third type of parameter card is the "LABEL" card. It is used to assign variable names to the variables in the analysis. Each name may be up to 8 characters long. The card is identified by the keyword name "LABEL" which must start in card column 1. The number in parentheses is the number of the variable that is to be labeled with the first name. It

is assumed that all succeeding labels are sequential. If more than one card is needed for the labels, the succeeding label cards should have the same format as the first except the number in parentheses should be the number of the variable that is to be labeled next. (Variable number is the same as the subscript of the variable as it is read into the input array X.) A comma or period should not be placed after the last label on the card. There can be any number of blank spaces on a card after the last label on that card. However, column 80 cannot be used (i.e., column 80 is not read). When making multiple runs on a set of data, the labels from the first run carry over to subsequent runs.

Example:

```
LABEL(1) = FIRST, CAPACIT, DIØDE, ..., TEMP

LABEL(12) = TWELFTH,THIRTEEN, etc.
```

The above cards would assign X(1) the name FIRST, X(2) the name CAPACIT and so on til X(13) would be named THIRTEEN. There is no limit to the number of label cards nor the number of names per card. A variable name may not be continued from one card to the next. Any spaces between the delimiting commas are taken as part of the label and are counted as characters in the name. All or none, or any portion of the variables may be labeled.

D. BACKWARD card, (required if the STEPWISE card is not used).

If the user desires a backward elimination solution for "best" subset selection, a "BACKWARD" card must be included among the parameters. The form is

```
BACKWARD,SIG= alphahat
```

where *alphahat* is replaced by the significance level the user wishes to use

for deleting variables from the model. Variables will be dropped from the model until only variables that have alphahat values less than or equal to *alphahat* are left in the model.

In you have only one independent variable, you must use the BACKWARD card instead of the STEPWISE card.

E. FORCE card (optional).

If the user wishes to keep certain variables in the regression model, regardless of their contribution to the model, he may do so by using the "FORCE" statement. The "FORCE" statement will keep specified variables in either the backward elimination or stepwise solution. The following is an example of the "FORCE" card:

```
FØRCE,3,6,8.
```

The above card indicates that variables 3, 6, and 8 are to be kept in the model regardless of their contribution to the model. The "FORCE" card must end with a period. The maximum number of variables that may be forced into the model is 10. Do not force all of the independent variables into the model. If the user desires to have all of the independent variables in the model, then he should use the BACKWARD option (with SIG=1.0).

F. OUTPUT card (required)

The "OUTPUT" card may have any of the following specifications separated by commas.

Keyword	Action
1. CORR	Print simple correlations among all variables specified in the model.
2. SSXP	Print the corrected sum of squares and cross-products matrix for all variables specified in the model.
3. INVERSE	Print the inverse correlation matrix at each step in the analysis for all independent variables used at that step.
4. STEPS	Print an analysis of variance table and the regression coefficient estimates for each step of

the backward elimination or stepwise variable selection procedure. If not specified, only the results for the final model will be printed.

5. RESIDUALS — Print the y, $\hat{y}$ and residual for each observation using the final regression model.
Caution: Do not request RESIDUALS if the number of observations is large. This could produce excessive, if not unwanted, output.

6. ALL — Options 1 through 5 will all be in effect.

Note: Any subset of these options may be requested but should follow the order of appearance given above. The following example shows how the "OUTPUT" card may be used.

```
ØUTPUT,CØRR,INVERSE,STEPS,RESIDUALS
```

G. PLOT RESIDUALS card, (optional).

The user may request that residuals be plotted on the line printer by inserting the "PLOT RESIDUALS" card. The plots produced are: each independent variable in the final model versus the dependent variable; residuals versus time; residuals versus $\hat{y}$'s; and residuals versus each of the independent variables in the model. Time is assumed to be the same as order of data input, i.e., observation 1 is assumed to be recorded first in time, observation 2 as second, etc. The "PLOT RESIDUALS" card should be as follows:

```
PLØT RESIDUALS
```

H. STEPWISE card, (required if the BACKWARD card is not used).

The "STEPWISE" card is used to specify that the program should find the "best" subset of the full model using the stepwise procedure. The user may specify a significance level for deleting a variable from the model. The program will continue to add variables to the model as long as it can find a variable whose regression coefficient is significantly different

from zero at the specified significance level for entering variables into the model. At each step, t-tests are computed on each of the regression coefficients of the variables in the model and their alpha hats computed. If any of the t statistics are not significant at the level specified for deleting a variable, the least significant variable is dropped. This process continues until all variables in the model are significant at the specified deletion level. The program then searches for a new variable to be added to the model and this cycle is repeated until no new variables can be found which are significant at the specified significance level.

The program can be used for computing a forward solution, i.e., adding variables in order of their contribution but not deleting any, by setting the deletion significance level to 1.0.

An example of the "STEPWISE" specification is given below:

```
STEPWISE,SIGIN=0.05,SIGØUT=0.10
```

The above card indicates variables are to be added to the model as long as variables that are significant at the $\alpha = .05$ level can be found. Variables that are already in the model whose significance level rises above .10, i.e., $\hat{\alpha} > .10$, will be deleted from the model. If the values for SIGIN and SIGOUT are not specified, they will each default to .05. Note: SIGOUT must always be at least as large as SIGIN to avoid an indefinite loop where a variable may be introduced and then immediately deleted. Recall that if there is only one independent variable, BACKWARD elimination procedure should be used rather than the STEPWISE procedure.

I. <u>MODEL</u> card, (required).

The model card indicates which variables are in the model and whether they are dependent or independent variables. Up to 15 dependent variables may be specified for any single run. A separate analysis is done for each dependent variable. The model may have up to 179 independent variables if there is only one dependent variable. The total number of variables both read in and generated using transformations may not exceed 180. Therefore, if there are 15 dependent variables specified as the model card, the

maximum number of independent variables is 165. The model may be continued for as many cards as necessary, just be sure that the continuation occurs before or after a + sign, i.e., do not allow the continuation to interrupt a variable subscript. Do not start model continuation cards in column 1.

The writing of the model card is done using variable subscript numbers. The first variable read in is variable 1 and thus has subscript number 1, etc. The writing of the model may be in any of the 3 forms that follow, where the dependent variables are immediately after the word "MODEL", separated by commas and the independent variables are to the right of the equal sign and separated by plus signs.

1. MODEL,1,3,2=4+5+12.

This method simply uses the variable number to indicate which variables are in the model and whether they are dependent or independent variables. For this particular example, variables 1, 3 and 2 are the dependent variables and 4, 5 and 12 are the independent variables. A separate analysis would be performed for each of the dependent variables.

2. MODEL,Y1,Y3,Y2=X4+X5+X12.

This specification results in the same model as the method described above. It allows the user to use more specific notation which may be helpful in clarifying the model in some instances.

3. MODEL,Y(1),Y(3),Y(2)=B(4)X(4)+B(5)X(5)+B(12)X(12).

This specification is equivalent to the two above methods but allows more specificity in describing the model by allowing a dummy indicator, B(i), for the regression coefficients.

4. MODEL,1,Y3,Y(2)=4+B(5)X(5)+X12.

The above methods may be mixed. Note that the above models are all terminated by periods. The period is not necessary on the "MODEL" card but results in slightly faster parameter processing by eliminating the need for the program to determine whether the model card has been continued or not. The program always fits the model with the intercept assumed to already be included in the model. There may be any number of blank spaces between the independent variables as long as each variable specification is separated by a "+" sign. However, there can be no blank spaces to the left of the equal sign.

J. PRESS card (optional).

When regression techniques are used to build a response surface, the possibility exists of producing two competing models. This situation could easily arise in stepwise regression as a new surface is developed each time a new significant variable is added. Although "statistical significance" is a necessary condition for adding a new variable in stepwise regression, it is not an end in itself as there exists the possibility of overfitting the data. For example, it is possible to obtain a good fit on a set of points by using a polynomial of high degree. However, in doing so, one can overfit the data and produce a spurious model which makes poor predictions.

To protect against overfit, the Predicted Error Sum of Squares (PRESS) criterion as given in Allen (1971) can be used to determine the adequacy of a prediction model. For a regression model containing k variables and constructed from n observations, PRESS is computed in the following manner. For i=1,2,...,n, the ith observation is deleted from the original set of n observations and then a regression model containing the original k variables is constructed from the remaining n-1 observations. With this new regression model, the value $\hat{Y}_k(i)$ is the estimate for the deleted observation Y_i. Then, PRESS is defined from the preceding predictions and the n original observations by

$$\text{PRESS}_k = \sum_{i=1} \left(Y_i - \hat{Y}_k(i)\right)^2.$$

The regression model having the smallest PRESS value is preferred when choosing between two competing models, as this is an indication of how well the basic pattern of the data has been fit versus an overfit or an underfit.

To obtain the PRESS values at each step of the analysis, merely add the PRESS card to the deck of parameter cards. Also, if the analysis produces 2 or more steps, a plot of PRESS values versus step number is automatically produced under the PRESS option to make relative comparisons easier.

Example:

PRESS

K. RANK REGRESSION card (optional).

To obtain a regression analysis on the ranks of the data instead of the raw data merely add the RANK REGRESSION card to the deck of parameter cards. If the user requests RESIDUALS on the OUTPUT card, then the residuals are given as a rank residual (i.e., the difference between the rank of Y and the predicted rank of Y). In addition, the predicted ranks of the Y's are used to interpolate in the original raw Y's to obtain the raw $\hat{Y}$. This is done at each step in the analysis and these data are used to compute a "normalized" R^2. This R^2 value is associated with the raw data instead of the ranked data and is computed by

$$R^2 = \frac{\sum_{i=1}^{n} (\hat{Y}_i - \bar{Y})^2}{\sum_{i=1}^{n} (\hat{Y}_i - \bar{Y})^2 + \sum_{i=1}^{n} (Y_i - \hat{Y}_i)^2}$$

This "normalized" value of R^2 varies between zero and one and will be close to one if the model being analyzed predicts the observed values adequately.

Note that the adjusted total sum of squares can be factored into three components: (1) sum of squares of regression, (2) sum of squares of error and (3) sum of cross-products. The formulas are as follows:

$$\sum_{i=1}^{n} (Y_i - \bar{Y})^2 = \sum_{i=1}^{n} (\hat{Y}_i - \bar{Y})^2 + \sum_{i=1}^{n} (Y_i - \hat{Y}_i)^2 + \sum_{i=1}^{n} (\hat{Y}_i - \bar{Y})(Y_i - \hat{Y}_i).$$

If the $\hat{Y}_i$ are the usual least squares predictions then this last term (the sum of cross-products) is easily shown to be zero. However, since the $\hat{Y}_i$ computed here by the rank regression technique are definitely not the least squares predictions, then this term is non-zero (can be positive or negative). Hence, if the "usual" R^2 = (sum of squares of regression)/(total sum of squares) were computed, it would be either increased or

decreased over the "normalized" R^2 described above depending on whether the sum of cross-products is negative or positive. Therefore, it is felt that the normalized R^2 is a more appropriate measure of fit than the usual R^2 in this situation.

The sum of cross-products can be quite large in absolute value, depending upon the structure of the observations. If the raw data have unusual spacings or exhibit a non-monotone relationship, then the interpolation scheme used to obtain the $\hat{Y}_i$'s will not be adequate and the resulting sum of cross-products will be large. Hence as a measure of the "ability (or inability) to interpolate in the data" a Coefficient of Interpolation is computed and given for each model analyzed. This coefficient is computed by

$$\left| 2* \left[\frac{\sum_{i=1}^{n} (\hat{Y}_i - \bar{Y})^2 + \sum_{i=1}^{n} (Y_i - \hat{Y})^2}{2* \sum_{i=1}^{n} (\hat{Y}_i - \bar{Y})^2 + \sum_{i=1}^{n} (Y_i - \hat{Y}_i)^2 + \sum_{i=1}^{n} (\hat{Y}_i - \bar{Y})(Y - \hat{Y}_i)} \right] -1 \right| .$$

Note that if the sum of cross-products is zero then this coefficient will be zero and that as the sum of cross-products departs from zero, the coefficient approaches one. Therefore as a matter of interpretation, a value of this coefficient that is "near zero" indicates that the interpolation is adequate. If the coefficient in "near one," then the user should examine the original observations for unusual spacings and/or non-monotone relations in the data (such as cyclic).

After the final model has been selected, then the raw residuals are given (when RESIDUALS are requested) as the difference between the raw Y and the raw $\hat{Y}$ (see Section V for how to save these values for further analysis). The reader is referred to Iman and Conover (1979) for a thorough discussion of the rank regression analysis.

When using RANK REGRESSION, there is a limit on the number of observations that can be handled. Let NV = the number of variables input to STEPWISE (see DATA card, page 2) and N = the number of observations. Then the following limits are in effect:

1. If NV = 2 (i.e., one independent and one dependent variable) then $N \leq 64980$.

L. WEIGHT card (optional).

The "WEIGHT" card allows the user to perform a weighted regression analysis of the data. The card has one argument as shown below:

```
WEIGHT=IWT
```

where IWT is the variable number where the weights will appear. (Variable number is the same as the subscript of the variable as it is read into the input array X.) If IWT is zero than weighted regression will not be done. However, the default value of IWT is zero so if the user does not want to use weighted regression then the "WEIGHT" card need not be included. Weighted regression may be done on either raw data or rank transformed data. The weights are normalized by the program so that the sum of the weight equals the number of observations. When the WEIGHT option is requested the equations on the previous pages with respect to PRESS and the R^2 on ranks are automatically adjusted to reflect these weights, as is the case for the normal equations used in the regression analysis.

The card is identified by the keyword "WEIGHT" which must begin in column one. There must be no blanks before or after the equal sign. A period to terminate the card is optional.

Example:

```
WEIGHT=3
```

In the example, weighted regression has been requested. The weights will appear in the data as variable number three.

M. DROP card (optional).

The "DROP" card allows the user to unconditionally drop observations from the regression analysis. (For information on how to conditionally drop observations, see the "Additional Data Transformations" section.) If the user knows in advance that some observations will need to be discarded, he may use the "DROP" card, indicating the number(s) of the observation(s) to be dropped. The card is identified by the keyword "DROP" which must

begin in column one. The only blank necessary is the one after the word "DROP." Commas are used to separate the observation numbers and a period is required after the last observation number. The observations being dropped need not appear in any special order. There must be no blanks preceding the period and there must be no blanks before or after the commas unless the card is to be continued. To continue the "DROP" card, at least one blank must follow the last comma and the continuation card must begin in column two. There is no limit to the number of continuation cards or to the number of observation drops per card.

Example:

```
DROP 9,6,1,5.
```

In this example, the first, fifth, sixth, and ninth observations will be dropped. If the user wishes to drop a large number of observations, it may be advantageous to use the trans subroutine discussed in the "Additional Data Transformations" section.

N. TRANSFORMATION card (optional).

The "TRANSFORMATION" card allows the user to define new variables or redefine existing variables via simple transformations of existing variables. (For more complicated variable transformations, see the "Additional Data Transformations" section.) The writing of the "TRANSFORMATION" card is done using variable subscript numbers. The card is identified by the keyword "TRANSFORMATION" which must begin in column one. The only blank necessary is the one after the word "TRANSFORMATION" and there must be no blanks within the transformation definitions. The variable being defined or redefined appears on the left hand side of an equal sign and the variables used to define or redefine it appear on the right hand side. There must be no blanks on either side of the equal sign and the variables need not appear in any special order. Stars are used to separate the variables on the right hand side and there must be no blanks on either side of any star in a transformation definition. There is no limit to the number of variables used in a transformation definition as long as it fits on one card. A minus sign in front of a

variable on the right hand side indicates that the reciprocal of that variable is wanted. There must be no blanks on either side of the minus sign. Commas are used to separate transformation definitions. There must be no blanks before or after a comma unless the card is to be continued. In such cases, the comma must be followed by at least one blank and the continuation card must begin in column two. Do not break up a transformation definition by continuing it on to the next card. There is no limit to the number of continuation cards or to the number of transformation definitions per card. A period is required after the last transformation definition and there must be no blanks preceding it.

Labels for the variables created by the "TRANSFORMATION" card are automatically generated using the characters of the transformation definition. When using the "TRANSFORMATION" card and/or the TRANS subroutine, the raw variables must be numbered from NV+1 to NV+NT, where NV is the number of input variables and NT is the number of new variables created via transformations. (For a complete description of the TRANS subroutine, see the "Additional Data Transformations" section.)

Example:

```
TRANSFORMATION    4 = 1*3, 2=-2, 8=6*-7
```

In this example, variable number four has been defined to be the product of variable number one and variable number three, variable number two has been redefined to be the reciprocal of itself, and variable number eight has been defined as variable number six times the reciprocal of variable number seven, that is variable eight is variable six divided by variable seven. Note that use of a minus sign on this card is used to denote a reciprocal.

2. If NV $\geq$ 3, then NV*N $\leq$ 175931.

When raw data are being analyzed, there essentially is no limit to the number of observations that can be processed. If the user has a data set falling outside of the above limits but would still like to do a rank regression, then it would be necessary for the user to rank the data outside of the STEPWISE program and then not use the RANK REGRESSION card. However, using this approach will not provide raw $\hat{Y}$'s within the STEPWISE program.

Example:

```
RANK REGRESSIØN
```

O. END OF PARAMETERS card (required).

The "END OF PARAMETERS" card is used to indicate the end of parameter card input.

Example:

```
END ØF PARAMETERS
```

P. FORMAT card (required, if data are read from cards).

A format card is required if data are to be read from cards. The format card may contain D, E, and F and (in some cases) I-numeric specifications and T and X positional specifications. The format must be enclosed in parentheses and conform to the syntax of the CDC FORTRAN IV variable format. The format may be continued on up to 10 cards.

Example:

```
(F2.1,F2.2,F4.0,1X,F1.0,3X,F2.0)
```

In most cases, the user will use the F and X format specifications. Listed below is a sample data card with data punched in columns 1 through 15.

```
145796043182461
```

The preceding format would read in the data card assigning the values listed below for the respective variables.

X(1)	X(2)	X(3)	X(4)	X(5)
1.4	0.57	9604.0	1.0	61.0

In the general case of Fw.d, w is the field width of the variable on the data card (the variables are read in sequentially from 1 to NV, where NV is the number variables to be read from the cards), i.e., the number of columns used to specify the values of the variable and d is the number of decimal places read in or the number of columns to the left of where the decimal point should be placed. The nX specification is used to skip columns on the data card where n is the number of columns to be skipped. If more than one data card is used to input the data from one observation, a slash in the format is used to specify that reading should continue on the next card.

Example:

(F2.1,6F2.2,8F5.3,7F4.1/10F5.2)

The above format card would cause the first 22 variables to be read from the first data card defining the first 22 values in the observation vector and the next 10 (23 to 32) to be read from the second data card completing that observation vector. There may be any number of cards per observation vector.

If more than one card is necessary for specifying the format, it should be continued for however many cards are necessary without any indication in the format itself that it is being continued. In other words, it should be written just as it would be written on one long card.

Q. END OF DATA trailer card (required when reading data from cards).

In order that the user does not have to specify the number of

observations he has in his data, the "END OF DATA" trailer card should follow the data. The trailer is necessary only when multiple sets of data are being processed in one computer run. The trailer card is used to separate the data of one data set from the parameters of the next. If only one set of data is being processed, there is no need for a trailer card, although it may be used. Do not use an END OF DATA card when DATDIS = 2.

Example:

```
END ØF DATA
```

IV. ADDITIONAL DATA TRANSFORMATIONS

It is possible, in addition to the simple transformations specified by the "TRANSFORMATION" card, to create complicated transformations by use of a user supplied subroutine called TRANS. The user may redefine user supplied variables or create new variables as transformations or combinations of the variables read in. When using the TRANS subroutine and/or the "TRANSFORMATION" card, the new variables must be numbered from NV+1 to NV+NT where NV is the number of input variables and NT is the number of new variables created via transformations. The simple variable transformations defined by the "TRANSFORMATION" card are always done first. The subroutine TRANS is then used to perform the more complicated transformations or to conditionally drop observations through use of IDROP as explained below. This subroutine may contain any legal FORTRAN statements. There are five variables made available to the subroutine via common block IMAN which the user may use to either make decisions while processing the input data or to modify the input status of an observation. These variables are NRAW, NTRANS, IDROP, IDUM, and IRANK and are discussed below. The general form of the subroutine is as follows:

```
SUBRØUTINE TRANS (X)
DIMENSIØN X(180)
CØMMØN/IMAN/NRAW,NTRANS,IDRØP,IDUM,IRANK
          :
    any transformations
          :
RETURN
END
```

The above cards are the minimum requirements for the subroutine. The subroutine requires a single argument, X, which is the array of input variables for a single observation, i.e., the observation ector X. The variables in common block IMAN are:

NRAW - is the current count of the raw observation being read in. It may be used to make decisions when making variable transformations based on the observation count or it may be used as a label if values are to be printed out. The value of NRAW must not be changed by the programmer.

NTRANS - is the current count of the transformed observation being processed. It will always be the same as NRAW except in the case of dropping observations. In this case, NTRANS will be the count of those observations which have not been dropped. The value of NTRANS must not be changed by the programmer.

IDROP - is used to indicate that the user wishes to drop an observation from the analysis. The value of IDROP is always zero upon entering the subroutine. When IDROP is set to any non-zero value, the observation vector being transformed at the time IDROP is set non-zero, will be dropped from the analysis.

IDUM - has no function within subroutine TRANS and the programmer should not change the value of IDUM within TRANS.

IRANK - is set within the main program and indicates when the user has requested RANK REGRESSION. When IRANK = 0, the usual regression analysis on the raw data has been requested. When IRANK = 1, RANK REGRESSION has been requested. The value of IRANK must not be changed by the programmer.

Example of using the TRANS subroutine:

```
      SUBRØUTINE TRANS(X)
      DIMENSIØN X(180)
      CØMMØN/IMAN/NRAW,NTRANS,IDRØP,IDUM,IRANK
      IF(NRAW.EQ.1)FAT=0.0
      IF (X(9).NE.0.0)GØ TO 4
      IDRØP=1
      RETURN
4     X(13)=(X(8)-X(7))/X(5)
      X(16)=X(11)/X(9)*100
      FAT=FAT + X(10)
      X(4)=ALØG(X(4))
      WRITE (3,101)(X(J),J=13,16),FAT,NTRANS
101   FORMAT (20X,5F8.3,3X,I5)
```

```
RETURN
END
```

In the above example, assume 12 variables are read in from cards (NV=12) according to the format specifications given by the user on the variable format card. The first time through, i.e., when NRAW=1, the variable FAT is set to zero. Each time through, X(9) is checked for being zero and in those cases when it is zero, the observation is dropped for all variables 1 through 16. If X(9) is non-zero, then X(13) and X(16) are created as transformations of existing variables. Also X(4) is transformed to its log (base e) and X(10) is accumulated in the location named FAT. Note that variables X(14) and X(15) are not defined in this subroutine. They may have been previously defined by use of a TRANSFORMATION card, thus the user is allowed complete freedom in creating transformed variables. Also, when X(9) is non-zero, the values of the new variables are printed out along with the cumulative sum of X(10) and the number of transformed observation (NTRANS).

An example of the use of IRANK is as follows. If the user is making multiple passes of the same data set and some of the passes involve RANK REGRESSION while the remaining passes involve the usual regression on the raw data, and furthermore some of the data is to be dropped from the raw analysis but not dropped from the rank analysis: then the programmer can make use of the passed value of IRANK to determine whether an observation should be dropped or whether it should remain in the analysis for that particular run. That is, if IRANK = 1, then you would jump over the statements that set IDRØP = 1. This prevents the deletion of the observation when RANK REGRESSION is requested.

V. OUTPUT OF THE REGRESSION COEFFICIENTS AND THE OBSERVATIONS WITH THEIR PREDICTED VALUES WHEN USING RANK REGRESSION

The user often has need of the regression coefficients from a particular fitted model. In order to accomodate these users, we have incorporated into STEPWISE the following option. These coefficients are automatically written onto TAPE 19 and the user can access these coefficients as described below.

For each distinct ANOVA table that STEPWISE produces in a run, a unique sequence number is attached to that ANOVA table and is printed at the bottom of the table. This sequence number begins with "101" and increments

in unit steps. Also the "number" of the variable that is referred to below is the number given to that variable when it is read in (or generated via transformations) and corresponds to the number of the variable that is used in the MODEL card.

The information written on TAPE 19 is arranged as follows. For each unique sequence number, a group of card images is written on TAPE 19 followed by two blank card images. Within this group of card images the first card contains the following information:

Columns 1 - 22	"Unique Sequence No. = "
Columns 23 - 27	Contains the unique sequence number for this group of cards in an I5 format.
Columns 28 - 66	"ANALYSIS FOR DEPENDENT VARIABLE"
Columns 67 - 69	Contains the number of the dependent variable used in this analysis in an I3 format.
Columns 70 - 72	"_ _ _"
Columns 73 - 80	Contains the label of the dependent variable (given in columns 67 - 69) in an A8 format.

The second card contains the TITLE (see Section II-A) in an 80 column alphameric format. All remaining cards within this group of cards are written in the following format with the explanation following:

FØRMAT(1X,I3,1X,I2,1X,4(I3,E15.8).

Columns 2 - 4	The number of independent variable included in the fitted model, exlcuding the constant term.
Columns 6 - 7	A card number that is unique within this group of cards.
Columns 9 - 80	Contains 4 sets of values, each set containing the number of the independent variable included in the fitted model followed by the estimate of the coefficicient.

Note: The first coefficient on the first card with a group always contains the constant term and its "number" is zero.

In order to access this information the user may use one of the following procedures. The user could catalog TAPE 19 as a permanent file and then read whatever information he desires from the permanent file. For example, the following sequence would catalog TAPE 19.

```
.
.
.
LDSET,PRESET=INF.
STEP,PL=40000.
CATALØG,TAPE19,THIS-FILE,CY=1,CN=DEFAULTPW.
EXIT.
EXIT.
```

If the user does not wish to catalog TAPE 19 as a permanent file but desires punched cards, then the following sequence would be appropriate.

```
.
.
.
LDSET,PRESET=INF.
STEP,PL=40000.
REWIND,TAPE19.
CØPYCF,TAPE19,PUNCH.
EXIT.
EXIT.
```

Or if the user wished to have a print-out of the card images along with the punched cards, then the following would suffice.

```
.
.
.
LDSET,PRESET=INF.
STEP,PL=40000.
REWIND,TAPE19
CØPYCF,TAPE19,PUNCH.
REWIND,TAPE19.
CØPYSBF,TAPE19,ØUTPUT.
EXIT.
EXIT.
```

As mentioned in Section II - K, the user can obtain the original observations and the predicted raw $\hat{Y}$'s when using RANK REGRESSION by retrieving them from TAPE 20.

When requesting RANK REGRESSION and PLOT RESIDUALS, the plots given will be the ranks of the data and the predicted ranks or rank residuals (whichever is appropriate). That is, none of the plots involve any of the raw data. Therefore, if the user wishes to see plots of the original observations versus the raw $\hat{Y}$'s, residuals, etc., then this must be done separate from the STEPWISE program.

To facilitate this type of further analysis, the STEPWISE program is set up to automatically write the original observations and the raw $\hat{Y}$'s on TAPE 20 as a binary file. If the user wants to save these for further

analysis, then TAPE 20 should be cataloged as a permanent file as explained in the previous part of this section.

The data file on TAPE 20 is created in the following manner. All records are written in unformated binary code. The first record is the TITLE card information, which is 80 columns of alphameric characters. The second record is an integer variable that gives the number of observations. All of the remaining records consist of two floating point numbers. The first is the original observed Y-value and the second is the predicted raw value of Y (i.e., the raw $\hat{Y}$). The number of records of pairs of data will be equal to the number of observations.

VI. DECK SETUP FOR STEPWISE

On the following page is an example of how to setup the cards for use in running STEPWISE. Cards 1 - 15, 22 - 32, and 46 - 64 all begin in card column one, while the Fortran statements in cards 16 - 21 start in column 7. Cards 35 - 45 are data cards in 5F6.0 format. Card number 1 is the JOB card and needs to be changed only to show the user's name and the box number. Note also that due to large core memory requirements of STEPWISE, that the extended core parameter must be set at 1200. Card number 2 must contain the user's social security number, division number and charge number. Cards 3 - 14 are control cards and will remain the same for all runs. The only exception being the use of TAPE 19 and TAPE 20 control cards that are explained in Section V. Cards 16 - 21 provide a dummy subroutine for transformations with any desired transformations (see page 17) following card number 20. Cards 23 - 32 are parameter cards (see pages 2-17). If the data is on cards, it will appear starting on card 33 with one vector of observations per card(s), followed by the END OF DATA card. Cards 47 - 53 reprocess the data with a backward solution. Note the absence of the LABEL card. The labels used in the first pass of the data will be used for this second analysis. Cards 54 - 64 reprocess the data with a stepwise solution and RANK regression.

Card No. DECK SETUP FOR STEPWISE

```
1   STE,T10,EC1200.
2   ACCØUNT,S509423684,D1223,G13,A0188000,RP,KUNC.
3   FTN,R=2,B=TRANZ.
4   REWIND,TRANZ.
5   ATTACH,ØBJECT,STEPWISE-RLI.
```

```
REWIND,ØBJECT.
CØPYL,ØBJECT,TRANZ,STEP.
ATTACH,SUBS,D1223-7600-LIBRARY.
ATTACH,FXDPLIB,FXDPLIB.
LIBRARY,SUBS,FXDPLIB.
LDSET,PRESET=INF.
STEP,PL=40000.
EXIT.
EXIT.
(END ØF RECØRD -- MULTI-PUNCH 7 8 9 in CØL 1).
      SUBRØUTINE TRANS(X)
      CØMMØN/IMAN/NRAW,NTRANS,IDRØP,IDUM,IRANK
      DIMENSIØN X(180)
C     ANY TRANSFØRMATIØNS GØ HERE
      RETURN
      END
(END ØF RECØRD -- MULTI-PUNCH 7 8 9 in CØL 1)
TITLE, EXAMPLE ØF STEPWISE ØPTION (DATA FRØM DRAPER AND SMITH,
p. 365 - 402)
DATA,5,0,1.
LABEL(1)=X1,X2,X3,X4,Y
MØDEL,5=1+2+3+4.
STEPWISE,SIGIN=.05,SIGØUT=.10
PRESS
ØUTPUT,ALL
PLØT RESIDUALS
END ØF PARAMETERS
(5F6.0)
   7.    26.     6.    60.     78.5
   1.    29.    15.    52.     74.3
  11.    56.     8.    20.    104.3
  11.    31.     8.    47.     87.6
   7.    52.     6.    33.     95.9
  11.    55.     9.    22.    109.2
   3.    71.    17.     6.    102.7
   1.    31.    22.    44.     72.5
   2.    54.    18.    22.     93.1
  21.    47.     4.    26.    115.9
   1.    40.    23.    34.     83.8
  11.    66.     9.    12.    113.3
  10.    68.     8.    12.    109.4
END ØF DATA
TITLE,EXAMPLE ØF BACKWARD ØPTIØN (DATA FRØM DRAPER & SMITH,
p. 364 - 402)
DATA,5,0,2.
MØDEL,5=1+2+3+4.
BACKWARD,SIG=.05
ØUTPUT,CØRR,STEPS,RESIDUALS
PRESS
END ØF PARAMETERS
TITLE,EXAMPLE ØF RANK REGRESSIØN WITH STEPWISE ØPTIØN (DATA FRØM
DRAPER/SMITH)
DATA,5,0,2.
LABEL(1)=RANK(X1),RANK(X2),RANK(X3),RANK(X4),RANK(Y)
```

```
57    MODEL,5=1+2+3+4.
58    STEPWISE,SIGIN=.05,SIGOUT=.10
59    PRESS
60    RANK REGRESSION
61    ØUTPUT,CØRR,STEPS,RESIDUALS
62    END ØF PARAMETERS
63    (END ØF INFØRMATIØN -- MULTI-PUNCH 6 7 8 9 IN CØL 1)
64    (END ØF INFØRMATIØN -- MULTI-PUNCH 6 7 8 9 IN CØL 1)
```

VII. DECK SETUP FOR STANDARDIZING DATA

Any of the variables may be standardized by using the Deck Setup on the following 2 pages. Remarks on page 22 explain the control cards 1 - 15. Cards 18 - 30 explain how to use the standardizing program. The user should note that card 43 must be made to match the data, and that the DATA parameter card must have a 2 as the third parameter (see page 2).

Card decks of this standardizing program are available from the authors of this report.

The standardizing program can be replaced by any program that the user would desire in terms of manipulating the input data set.

```
Card
 No.        DECK SETUP FOR STANDARDIZING DATA BEFORE RUNNING STEPWISE

 1    STE,T10,EC1200.
 2    ACCØUNT,S509423684,D1223,G13,A0188000,RP,KUNC.
 3    FTN,R=2,B=STAN.
 4    LDSET,PRESET=ZERØ.
 5    STAN.
 6    FTN,R=2,B=TRANZ.
 7    REWIND,TRANZ.
 8    ATTACH,ØBJECT,STEPWISE-RLI.
 9    REWIND,ØBJECT.
10    CØPYL,ØBJECT,TRANZ,STEP.
11    ATTACH,SUBS,D1223-7600-LIBRARY.
12    ATTACH,FXDPLIB,FXDPLIB.
13    LIBRARY,SUBS,FXDPLIB.
14    LDSET,PRESET=INF.
15    STEP,PL=40000.
16    (END ØF RECØRD -- MULTI-PUNCH 7 8 9 IN CØL 1)
17    C
18    C THIS PRØGRAM WILL STANDARDIZE ANY ØF THE INPUT VARIABLES.  FØR
19    C STANDARDIZING A CARD MUST BE PLACED IN FRØNT ØF THE DATA WITH THE
20    C NUMBER ØF VARIABLES (NV) RIGHT JUSTIFIED IN CØLUMNS 1 - 3.  THE
21    C REST ØF THIS CARD AND PØSSIBLY CØNTINUING ØNTØ A SECØND CARD DE-
22    C PENDING ON THE NUMBER OF VARIABLES) WILL BE FILLED WITH EITHER ZERØS
23    C ØR ØNES.  A 1 PUNCHED IN CØLUMN 4 WILL INDICATE THAT VARIABLE NUM-
24    BER ØNE IS TO BE STANDARDIZED, WHILE A 0 PUNCHED IN THIS CØLUMN WILL
25    INDICATE THAT VARIABLE NUMBER 1 IS NØT TØ BE STANDARDIZED.  VAR-
26    IABLE NUMBER 2 IS LIKEWISE FLAGGED IN CØLUMN 5, VARIABLE NUMBER 3 IN
```

```
27    CØLUMN 6, ETC.
28    C
29    C THE FIRST DIMENSIØN ØF THE DATA MATRIX SHØULD BE AT LEAST ØNE
30    C GREATER THAN THE NUMBER ØF ØBSERVATIØNS.  WHILE THE SECØND DIMEN-
31    C SIØN MUST BE AT LEAST EQUAL TØ NV.
32    C
33          PRØGRAM STAND(INPUT,ØUTPUT,TAPE10, TAPE 5 = INPUT)
34          DIMENSIØN IFLAG(101),DATA(10000,9)
35          READ(5,105)NV,(IFLAG(J),J=1,NV)
36      105 FØRMAT(I3,77I1/24I1)
37          NV1 = NV - 1
38    C
39    C     READ IN DATA.  FØRMAT 100 MUST CØNFORM TØ USERS DATA.
40    C
41          I = 1
42        1 READ(5,100)(DATA(I,J),J=1,NV)
43          IF(EOF(5))2,3
44      100 FØRMAT(2X,7F10,5,4XF4.0)
45        3 CØNTINUE
46          I = I + 1
47          GØ TØ 1
48        2 I = I - 1
49          PRINT 102
50      102 FØRMAT(*1MEANS AND ST. DEVS. ØF VARIABLES THAT WERE STANDARD
51         1IZED*,//,* VAR NØ.    MEAN        ST.DEV.*)
52          DØ 4 K = 1,NV
53          IF(IFLAG(K).EQ.0) GO TO 8
54          XB = 0.
55          SD = 0.
56          DØ 5 J = 1,I
57          XB = XB + DATA(J,K)
58        5 SD = SD + DATA(J,K)**2
59          XB = XB/I
60          SD = SQRT((SD - I*XB**2)/(I-1))
61          DØ 6 J = 1,I
62        6 DATA(J,K) = (DATA(J,K) - XB)/SD
63          PRINT 101,K,XB,SD
64      101 FØRMAT(*0*,I4,2F12.5)
65          GØ TØ 4
66        8 PRINT 106,K
67      106 FØRMAT(*0*,I4,*    THIS VARIABLE WAS NØT STANDARDIZED*)
68        4 CØNTINUE
69          DØ 7 J = 1,I
70        7 WRITE (10)(DATA(J,K),K=1,NV)
71          ENDFILE 10
72          REWIND 10
73          CALL EXIT
74          END
75    (END ØF RECØRD -- MULTI-PUNCH 7 8 9 IN CØL 1)
76
77
78    PARAMETER CARD(S) FØR STANDARDIZING VARIABLES GØES HERE
79
80    FØLLØWED BY DATA CARDS
```

```
81
82
83    (END ØF RECØRD -- MULTI-PUNCH 7 8 9 IN COL 1)
84          SUBRØUTINE TRANS(X)
85          CØMMØN/IMAN/NRAW,NTRANS,IDRØP,IDUM,IRANK
86          DIMENSIØN X(180)
87    C     ANY TRANSFØRMATIØNS GØ HERE
88          RETURN
89          END
90    (END OF RECØRD -- MULTI-PUNCH 7 8 9 IN CØL 1)
91
92          PARAMETER CARDS GØ HERE
93
94          DATA CARD MUST HAVE A 2 AS THE THIRD PARAMETER --
95
96          THIS INDICATES DATA IS ØN TAPE 10(DISK)
97
98    (END ØF INFORMATIØN -- MULTI-PUNCH 6 7 8 9 IN CØL 1)
99    (END ØF INFØRMATIØN -- MULTI-PUNCH 6 7 8 9 IN CØL 1)
```

VIII. BIBLIOGRAPHY

Allen, D. M. (1971). The Prediction Sum of Squares as a Criterion for Selecting Predictor Variables. Technical Report 23, University of Kentucky, Department of Statistics.

Iman, R. L. (1976). STEPWISE Regression. Technical Report, SAND76-0364, Sandia Laboratories, Albuquerque, NM.

Iman, R. L. and Conover, W. J. (1979). The Use of the Rank Transform in Regression. Technometrics, 21, 499-509.

IX. EXAMPLE

```
SANDIA LABORATORIES <> STEPWISE REGRESSION PROGRAM <> COURTESY OF DEPT. OF STATISTICS - KANSAS STATE UNIVERSITY
TITLE,EXAMPLE OF STEPWISE OPTION (DATA FROM DRAPER AND SMITH, P. 365-402)
DATA,5,0,1.                                                                    (STAT CONTROL CARD)
                    INPUT CHECK OF PARAMETERS
                    NUMBER OF VARIABLES READ IN =    5
                    NO. OF TRANSFORMED VARIABLES =    0
                    DATA DISPOSITION IS  1
LABEL(1)=X1,X2,X3,X4,Y                                                         (STAT CONTROL CARD)
MODEL,5=1+2+3+4.                                                               (STAT CONTROL CARD)
STEPWISE,SIGIN=.05,SIGOUT=.10                                                  (STAT CONTROL CARD)
PRESS                                                                          (STAT CONTROL CARD)
OUTPUT,ALL.                                                                    (STAT CONTROL CARD)
PLOT RESIDUALS                                                                 (STAT CONTROL CARD)
END OF PARAMETERS                                                              (STAT CONTROL CARD)
FORMAT    (5F6.0)
```

TITLE APPEARS ON EACH PAGE

5 INPUT VARIABLES

NO. TRANSFORMED VARIABLES

DATA WILL BE READ FROM CARDS AND SAVED ON DISK

5 IS THE DEPENDENT VARIABLE; 1,2,3,4 ARE INDEPENDENT VARIABLES

SIGNIFICANCE LEVEL FOR ENTERING VARIABLES

CALCULATE AND PLOT PRESS VALUES

OUTPUT: SUMS OF SQUARES MATRIX, CORRELATION MATRIX, INVERSE OF CORRELATION MATRIX, STEPS IN REGRESSION, AND RESIDUALS

OUTPUT LINE PRINTER PLOTS

```
TITLE,EXAMPLE OF STEPWISE OPTION (DATA FROM DRAPER AND SMITH, P. 365-402)          PAGE  1
SANDIA LABORATORIES <><>  STEPWISE  REGRESSION  <><> FROM KANSAS STATE UNIVERSITY
                    INPUT CHECK ON DATA
                    FIRST OBSERVATION
X1          X2          X3          X4          Y
  1           2           3           4           5
7.000000   26.000000   6.000000   60.000000   78.500000
                    NO. RAW DATA INPUT =   13
                    NO. TRANSFORMED OBSERVATIONS =   13
                    NO. OF OBSERVATIONS DROPPED =    0
```

TITLE,EXAMPLE OF STEPWISE OPTION (DATA FROM DRAPER AND SMITH, P. 365-402) PAGE 2

SANDIA LABORATORIES <><> STEPWISE REGRESSION <><> FROM KANSAS STATE UNIVERSITY

VARIABLE NAME	VARIABLE NUMBER	MEAN	VARIANCE	STD. DEV.	STD. ERR.	C.V.
X1	1	7.46154	34.6026	5.88239	1.63148	78.8+
X2	2	48.1538	242.141	15.5609	4.31581	32.31
X3	3	11.7692	41.0256	6.40513	1.77648	54.42
X4	4	30.0000	280.167	16.7382	4.64234	55.79
Y	5	95.4231	236.314	15.0437	4.17238	15.77

$S.E. = \frac{ST.\ DEV.}{\sqrt{n}}$ $C.V. = \frac{ST.\ DEV.}{\bar{X}} * 100\%$

13 OBSERVATIONS

TITLE,EXAMPLE OF STEPWISE OPTION (DATA FROM DRAPER AND SMITH, P. 365-402) PAGE 3

SANDIA LABORATORIES <><> STEPWISE REGRESSION <><> FROM KANSAS STATE UNIVERSITY

SUM OF SQUARES MATRIX

X1	1	4.152E+02				
X2	2	2.511E+02	2.906E+03			
X3	3	-3.726E+02	-1.665E+02	4.923E+02		
X4	4	-2.902E+02	-3.041E+03	7.800E+01	3.362E+03	
Y	5	7.760E+02	2.233E+03	-6.182E+02	-2.482E+03	2.716E+03
	NO.	1	2	3	4	5
NAME		X1	X2	X3	X4	Y

TITLE,EXAMPLE OF STEPWISE OPTION (DATA FROM DRAPER AND SMITH, P. 365-402) PAGE 4

SANDIA LABORATORIES <><> STEPWISE REGRESSION <><> FROM KANSAS STATE UNIVERSITY

CORRELATION MATRIX → FOR ALL THE INPUT DATA

X1	1	1.0000				
X2	2	.2286	1.0000			
X3	3	-.8241	-.1392	1.0000		
X4	4	-.2454	-.9730	.0295	1.0000	
Y	5	.7307	.8163	-.5347	-.8213	1.0000
	NO.	1	2	3	4	5
NAME		X1	X2	X3	X4	Y

HIGHEST CORR. ⇒ X_4 IS FIRST INDEPENDENT VARIABLE ADDED TO THE MODEL

$$r_{25.4} = \frac{.8163-(-.9730)(-.8213)}{\sqrt{(1-(-.9730)^2)(1-(-.8213)^2)}} = .1304$$

$$r_{35.4} = \frac{-.5347-(.0295)(-.8213)}{\sqrt{(1-(.0295)^2)(1-(-.8213)^2)}} = -.8952$$

PARTIAL CORRELATIONS AFTER X_4 IS ADDED TO MODEL

$$r_{15.4} = \frac{.7307-(-.2454)(-.8213)}{\sqrt{(1-(-.2454)^2)(1-(-.8213)^2)}} = .9568$$

= HIGHEST PARTIAL CORR. ⇒ X_1 WILL BE THE SECOND VARIABLE ADDED AFTER X_4

TITLE,EXAMPLE OF STEPWISE OPTION (DATA FROM DRAPER AND SMITH, P.369-402) PAGE 5

SANDIA LABORATORIES <><> STEPWISE REGRESSION <><> FROM KANSAS STATE UNIVERSITY

INVERSE OF CORR MATRIX → FOR INDEPENDENT VARIABLE SELECTED IN STEP 1.

X4 4 1.000E+00

NO. 4

NAME X4

TITLE,EXAMPLE OF STEPWISE OPTION (DATA FROM DRAPER AND SMITH, P. 369-402) PAGE 5

SANDIA LABORATORIES <><> STEPWISE REGRESSION <><> FROM KANSAS STATE UNIVERSITY

ADV TABLE
ANALYSIS OF REGRESSION FOR VARIABLE 5 --- → THE DEPENDENT VARIABLE IS INPUT VARIABLE 5
(TABLE 1)

SOURCE	D.F.	SS	MS	F	SIGNIFICANCE
REGRESSION	1	1831.8952	1831.8952	22.798520	.0006
RESIDUAL	11	883.86692	80.351538		
TOTAL	12	2715.7621			

$F = t^2$ AT STEP 1

R**2 IS .67454
INTERCEPT IS 117.56793
STANDARD ERROR OF INTERCEPT IS 5.26221

VARIABLE NUMBER	VARIABLE NAME	REGRESSION COEFFICIENTS	STANDARDIZED REGRESSION COEFFICIENTS	PARTIAL SSQ	T-TEST VALUES } 11 D.F.	R**2 DELETES	ALPHA FATS
4	X4	-.73816181	-.821365	1831.8952	-4.7748	0.0000	.0011

USED FOR TESTING $H_0: \beta_4 = 0$

UNIQUE SEQUENCE NUMBER FOR THIS ANOVA = 101

THIS IS THE 1ST ANOVA TABLE WITHIN THIS RUN OF STEPWISE

PRESS IS .119E+04

UNIT FREE BETA, I.E. $\beta_4^* = \hat{\beta}_4 \sqrt{\frac{\sum (x_{4i} - \bar{x}_4)^2}{\sum (y_i - \bar{y})^2}}$

--- THE β^*'S CAN BE USED FOR ORDERING THE RELATIVE IMPORTANCE WHEREAS THE USUAL $\hat{\beta}$'S CANNOT BE USED FOR ORDERING

REGRESSION EQUATION: $\hat{Y} = 117.56793 - .73816181 * X_4$

THE CONSTANT TERM

TITLE,EXAMPLE OF STEPWISE OPTION (DATA FROM DRAPER AND SMITH, P. 365-402) PAGE 2

SANDIA LABORATORIES <><> STEPWISE REGRESSION <><> FROM KANSAS STATE UNIVERSITY

INVERSE OF CORR MATRIX → FOR INDEPENDENT VARIABLES SELECTED IN STEP 2

X1	1	1.064E+00	
X4	4	2.612E-01	1.064E+00

NO.	1	4
NAME	X1	X4

TITLE,EXAMPLE OF STEPWISE OPTION (DATA FROM DRAPER AND SMITH, P. 365-402) PAGE 3

SANDIA LABORATORIES <><> STEPWISE REGRESSION <><> FROM KANSAS STATE UNIVERSITY

ADV TABLE
ANALYSIS OF REGRESSION FOR VARIABLE 5---Y
(TABLE 1)

SOURCE	D.F.	SS	MS	F	SIGNIFICANCE
REGRESSION	2	2641.0610	1320.5305	176.62696	.0000
RESIDUAL	10	74.762112	7.4762112		
TOTAL	12	2715.7631			

R**2 IS .97247
INTERCEPT IS 103.09738
STANDARD ERROR OF INTERCEPT IS 2.12398

VARIABLE NUMBER	VARIABLE NAME	REGRESSION COEFFICIENTS	STANDARDIZED REGRESSION COEFFICIENTS	PARTIAL SSQ	T-TEST VALUES	R**2 DELETES	ALPHA MATS
1	X1	1.4399583	.563052	809.1648	10.4031	.6745	.0001
4	X4	-.61395363	-.683107	1190.9246	-12.6212	.5339	.0000

THIS IS THE VALUE THAT R^2 WILL DECREASE TO IF X_4 IS NOW DELETED FROM THE MODEL, I.E.

SIGNIFICANCE LEVEL HAD TO BE < SIGIN = .05 FOR THE VARIABLE TO ENTER THE MODEL

UNIQUE SEQUENCE NUMBER FOR THIS ANOVA = 102

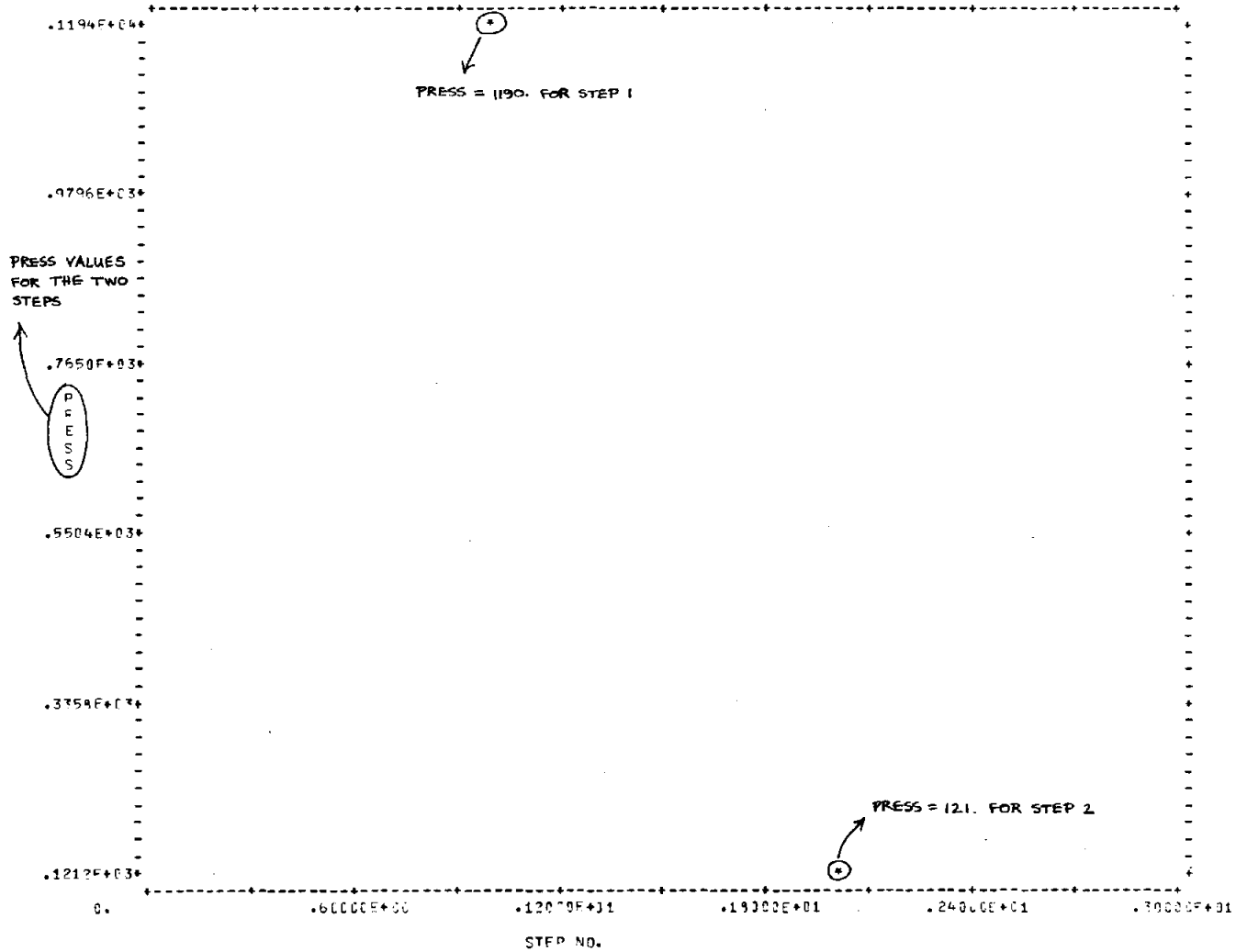
PRESS IS 121.

$$.5339 = .97247 - \frac{1190.9246}{2715.7631}$$

FINAL REGRESSION EQUATION : $\hat{Y} = 103.09738 + 1.4399583 * X_1 - .61395363 * X_4$

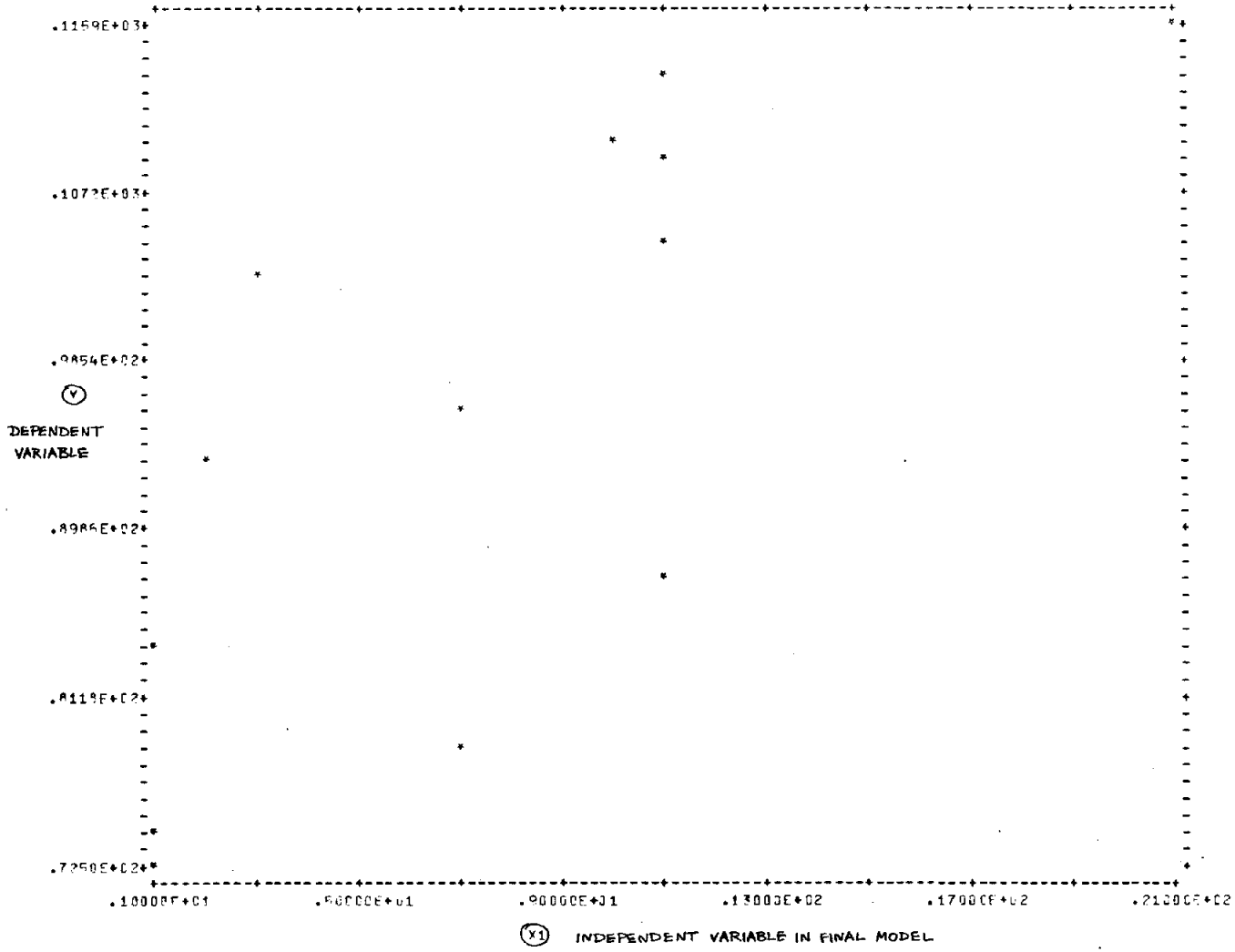
TITLE,EXAMPLE OF STEPWISE OPTION (DATA FROM DRAPER AND SMITH, P. 389-402) PAGE 3

SANDIA LABORATORIES <><> STEPWISE REGRESSION <><> FROM KANSAS STATE UNIVERSITY



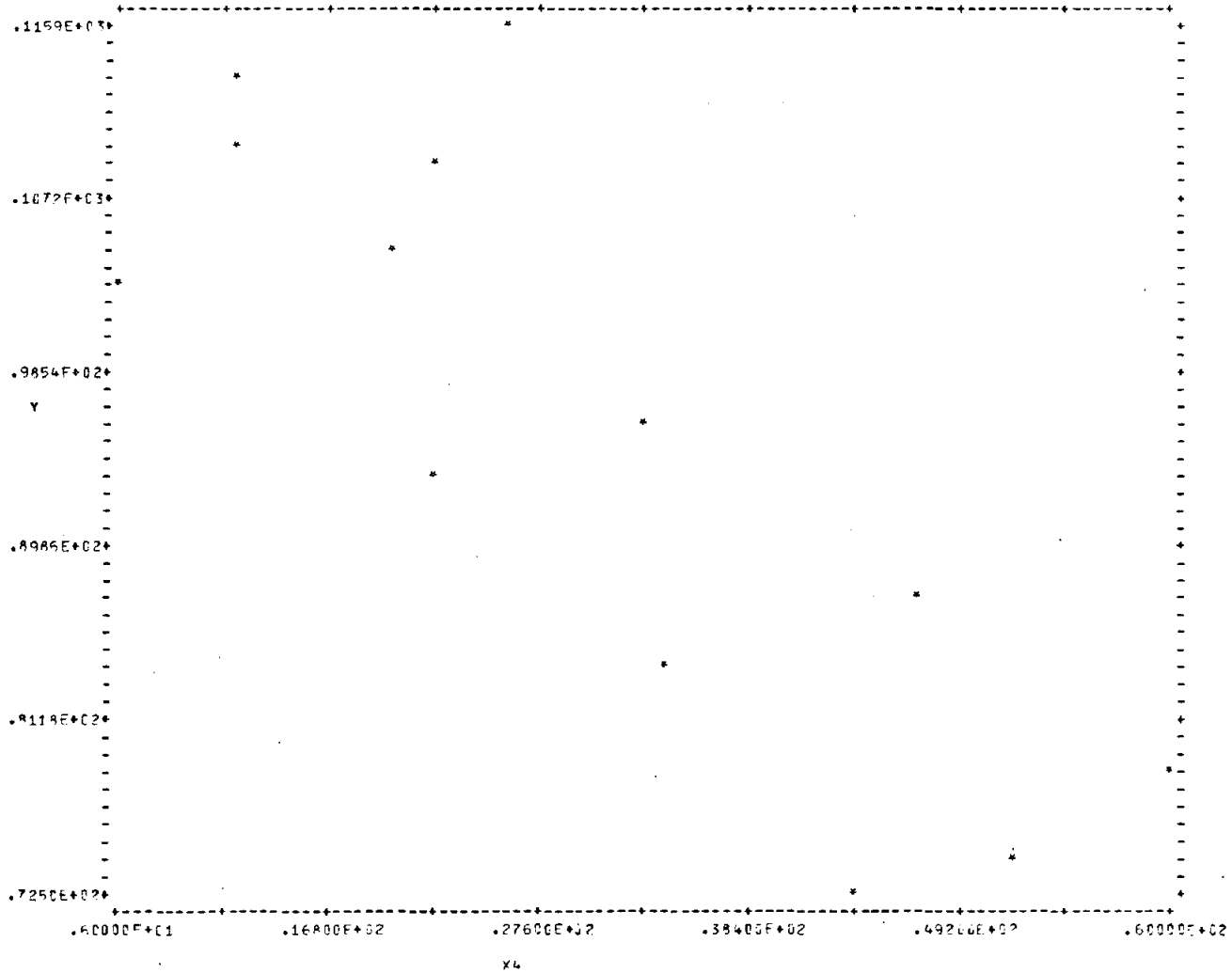
TITLE,EXAMPLE OF STEPWISE OPTION (DATA FROM DRAPER AND SMITH, P. 385-402) PAGE 13

SANDIA LABORATORIES <><> STEPWISE REGRESSION <><> FROM KANSAS STATE UNIVERSITY



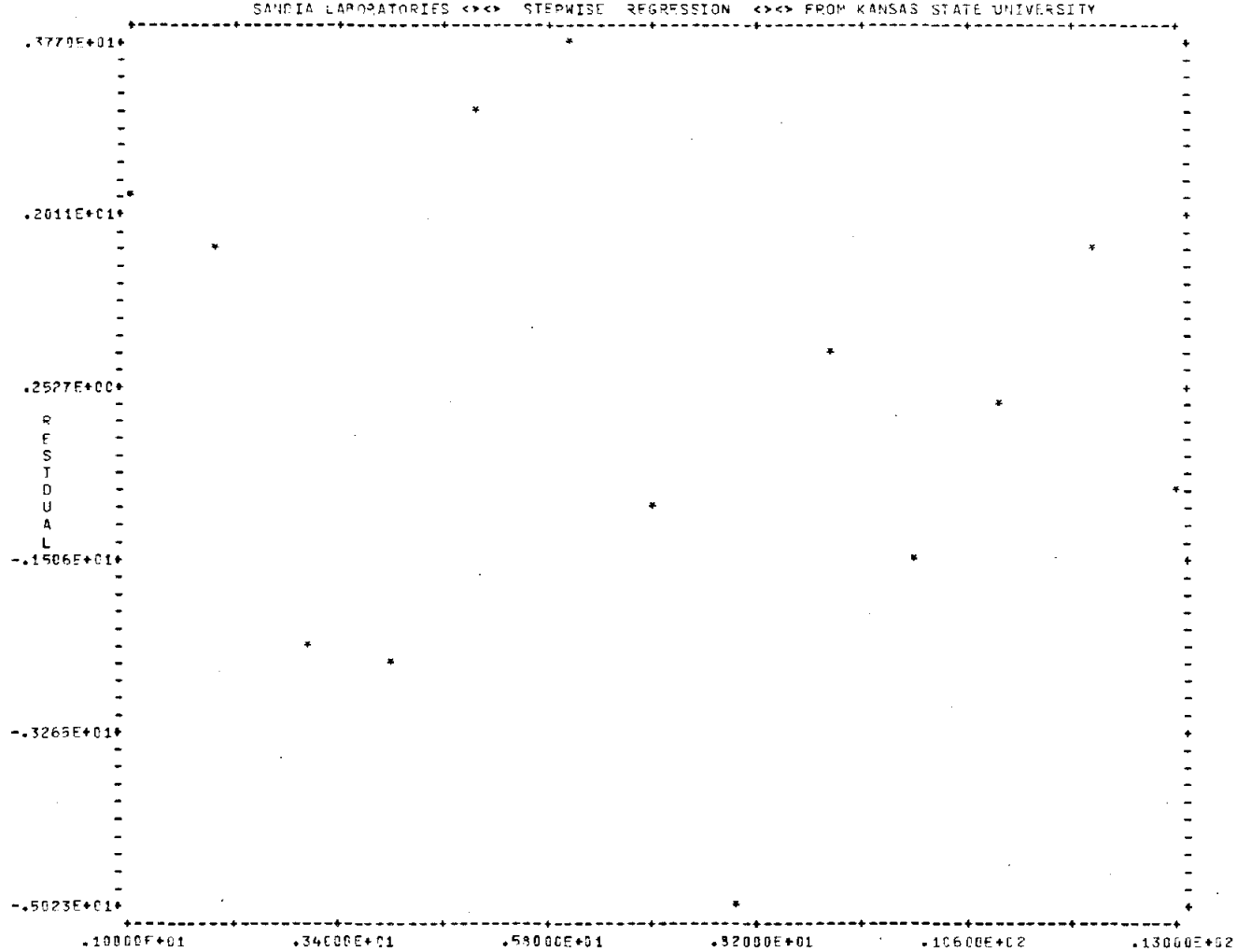
TITLE,EXAMPLE OF STEPWISE OPTION (DATA FROM DRAPER AND SMITH, P. 385-402) PAGE 11

SANDIA LABORATORIES <><> STEPWISE REGRESSION <><> FROM KANSAS STATE UNIVERSITY



TITLE,EXAMPLE OF STEPWISE OPTION (DATA FROM DRAPER AND SMITH, P. 365-402) PAGE 12

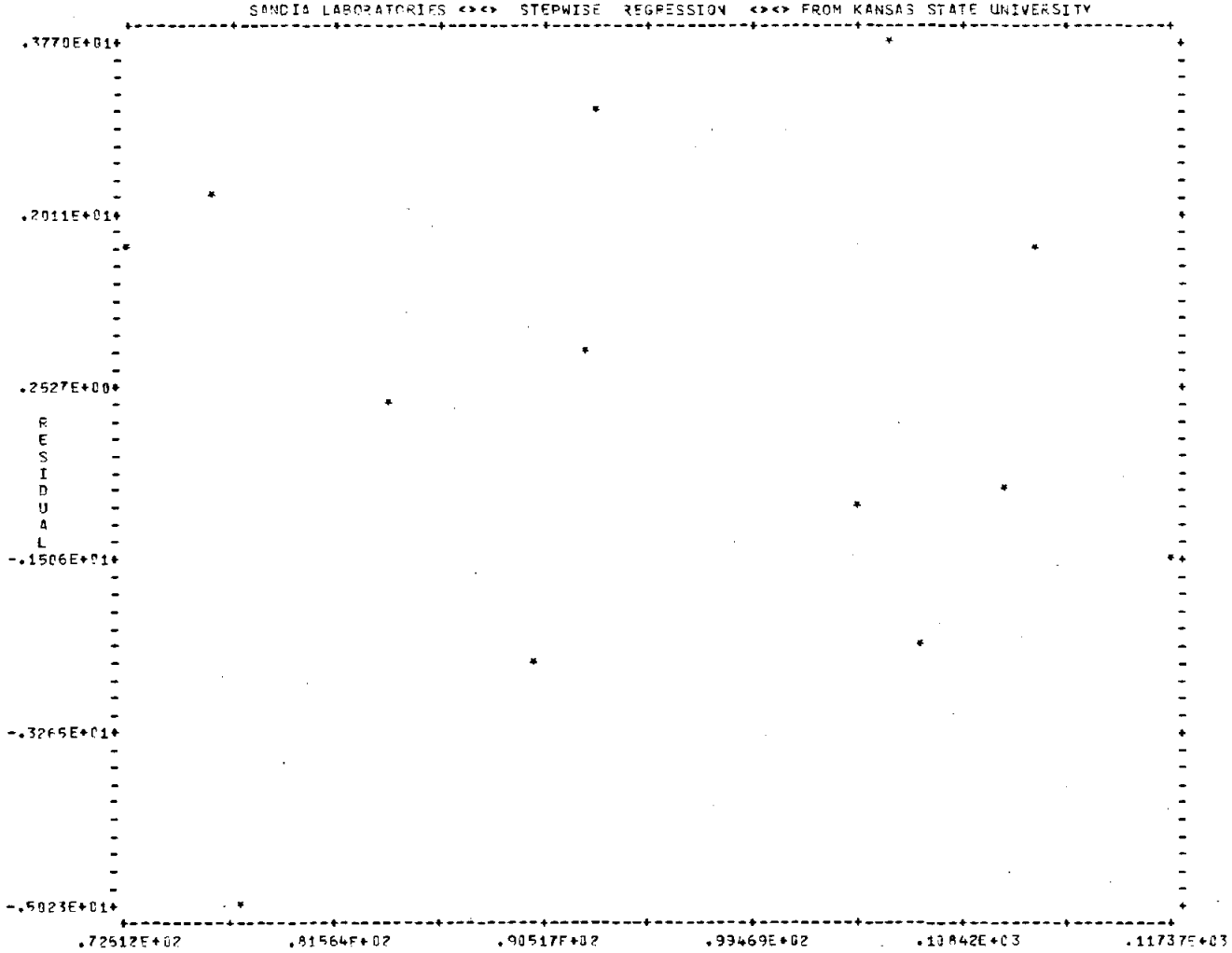
SANDIA LABORATORIES <><> STEPWISE REGRESSION <><> FROM KANSAS STATE UNIVERSITY



ORDER IN WHICH DATA WAS READ IN

TITLE,EXAMPLE OF STEPWISE OPTION (DATA FROM DRAPER AND SMITH, P. 365-402)

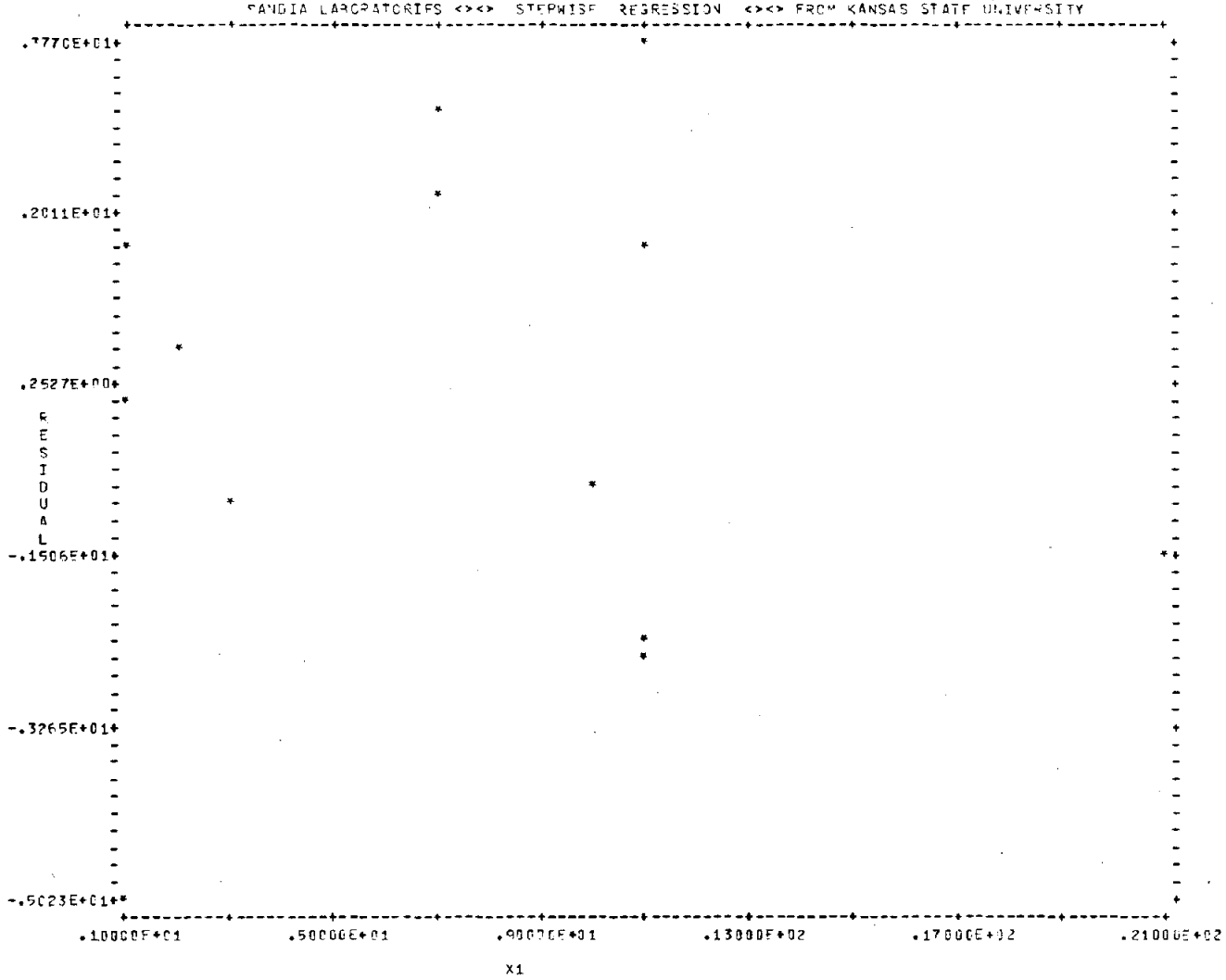
SANDIA LABORATORIES <><> STEPWISE REGRESSION <><> FROM KANSAS STATE UNIVERSITY



PREDICTIONS USING FINAL REGRESSION EQUATION

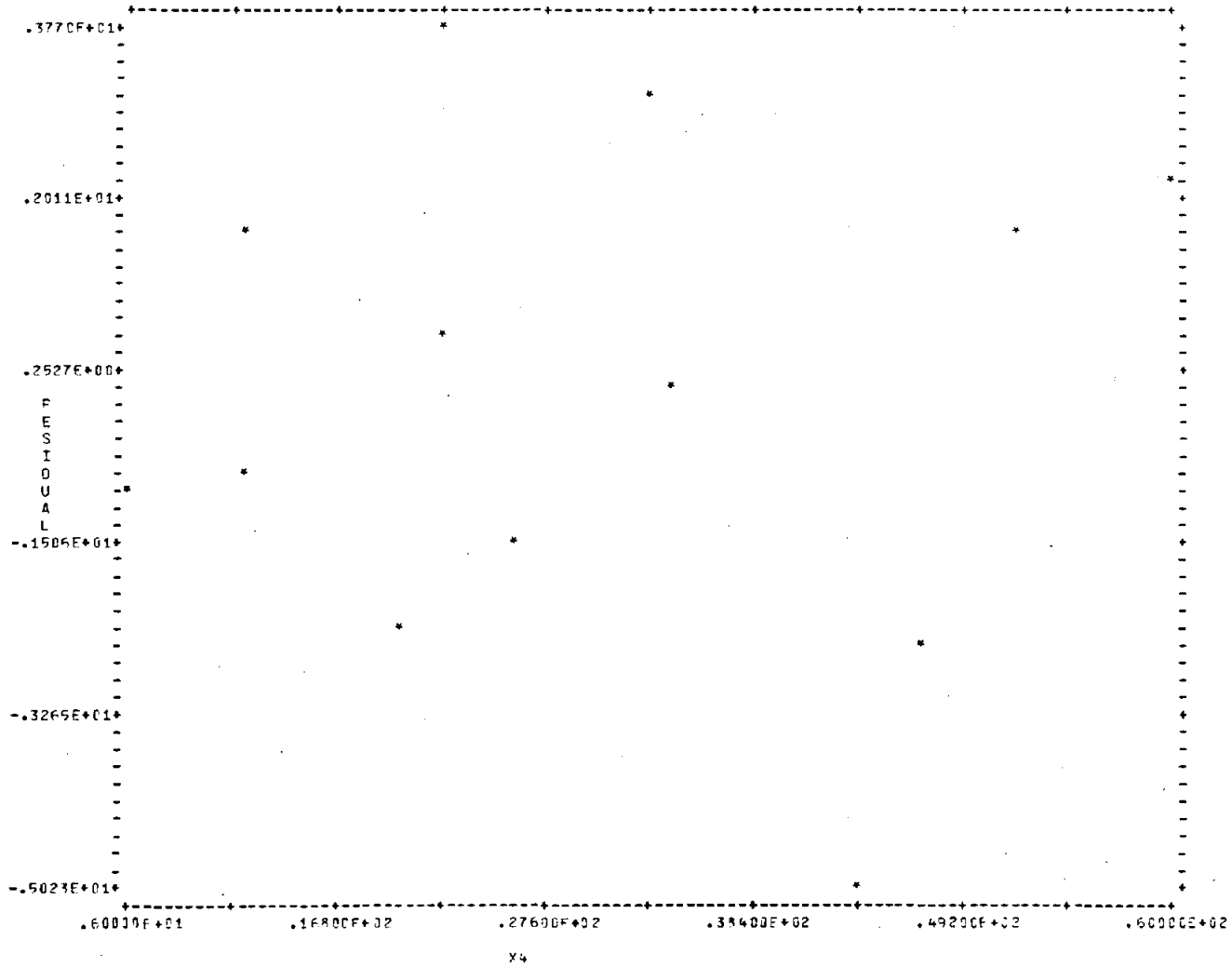
TITLE.EXAMPLE OF STEPWISE OPTION (DATA FROM DRAPER AND SMITH, P. 395-402) PAGE 14

SANDIA LABORATORIES <>< STEPWISE REGRESSION <>< FROM KANSAS STATE UNIVERSITY



TITLE,EXAMPLE OF STEPWISE OPTION (DATA FROM DRAPER AND SMITH, P. 365-412) PAGE 15

SANDIA LABORATORIES <><> STEPWISE REGRESSION <><> FROM KANSAS STATE UNIVERSITY



TITLE,EXAMPLE OF STEPWISE OPTION (DATA FROM DRAPER AND SMITH, P. 365-402) PAGE 15

SANDIA LABORATORIES <><> STEPWISE REGRESSION <><> FROM KANSAS STATE UNIVERSITY

TABLE OF RESIDUALS FOR VARIABLES 5----Y

TIME	OBSERVED VALUE	PREDICTED VALUE	RESIDUAL
1	78.6000	76.3339	2.16613
2	74.3000	72.6118	1.68825
3	104.300	106.658	-2.35785
4	87.6000	90.0811	-2.48110
5	95.9000	92.9166	2.98338
6	109.200	105.430	3.77006
7	102.700	103.734	-1.03353
8	72.5000	77.5234	-5.02338
9	93.1000	92.4703	.629682
10	115.900	117.374	-1.47371
11	83.8000	83.5629	.137083
12	113.300	111.569	1.73052
13	109.400	110.130	-.729521

Y $\hat{Y}$ $(Y-\hat{Y})$

END OF ANALYSIS FOR 1ST PASS ON THE DATA

SANDIA LABORATORIES <> STEPWISE REGRESSION PROGRAM <> COURTESY OF DEPT. OF STATISTICS - KANSAS STATE UNIVERSITY

TITLE,EXAMPLE OF BACKWARD OPTION (DATA FROM DRAPER AND SMITH, P. 365-402)

DATA,5,0,2. (STAT CONTROL CARD)

DATA WAS SAVED ON DISK ON THE FIRST PASS.

INPUT CHECK OF PARAMETERS

NUMBER OF VARIABLES READ IN = 5

NO. OF TRANSFORMED VARIABLES = 0

DATA DISPOSITION IS 2

MODEL,5=1+2+3+4. (STAT CONTROL CARD)

NOTE THE LABELS CARD IS MISSING: LABELS WILL CARRY OVER FROM 1ST RUN.

BACKWARD,SIG=.05 → LEVEL FOR DELETING VARIABLES (STAT CONTROL CARD)

OUTPUT,CORR,STEPS,RESIDUALS. → NO INVERSE, NO SUMS OF SQUARES MATRIX (STAT CONTROL CARD)

PRESS (STAT CONTROL CARD)

END OF PARAMETERS (STAT CONTROL CARD)

TITLE,EXAMPLE OF BACKWARD OPTION (DATA FROM DRAPER AND SMITH, P. 365-402) PAGE 1

SANDIA LABORATORIES <><> STEPWISE REGRESSION <><> FROM KANSAS STATE UNIVERSITY

INPUT CHECK ON DATA

FIRST OBSERVATION

X1	X2	X3	X4	Y
1	2	3	4	5
7.0000000	26.000000	6.0000000	60.000000	78.500000

NO. RAW DATA INPUT = 13

NO. TRANSFORMED OBSERVATIONS = 13

NO. OF OBSERVATIONS DROPPED = 0

TITLE,EXAMPLE OF BACKWARD OPTION (DATA FROM DRAPER AND SMITH, P. 365-402) PAGE 2

SANDIA LABORATORIES <><> STEPWISE REGRESSION <><> FROM KANSAS STATE UNIVERSITY

VARIABLE NAME	VARIABLE NUMBER	MEAN	VARIANCE	STD. DEV.	STD. ERR.	C.V.
X1	1	7.46154	34.6026	5.88239	1.63148	78.84
X2	2	48.1538	242.141	15.5609	4.31581	32.31
X3	3	11.7692	41.0256	6.40513	1.77646	54.42
X4	4	30.0000	280.167	16.7332	4.64234	55.79
Y	5	95.4231	226.314	15.0437	4.17238	15.77

13 OBSERVATIONS

TITLE,EXAMPLE OF BACKWARD OPTION (DATA FROM DRAPER AND SMITH, P. 365-402) PAGE 3

SANDIA LABORATORIES <><> STEPWISE REGRESSION <><> FROM KANSAS STATE UNIVERSITY

CORRELATION MATRIX

X1	1	1.0000				
X2	2	.2286	1.0000			
X3	3	-.8241	-.1392	1.0000		
X4	4	-.2454	-.9730	.0295	1.0000	
Y	5	.7307	.8163	-.5347	-.8213	1.0000
	NO.	1	2	3	4	5
NAME		X1	X2	X3	X4	Y

ROUNDING ERROR IN ROW 3

FLAG INDICATING HIGHLY CORRELATED INPUT VARIABLES.
IF MATRIX IS SINGULAR A MESSAGE WILL BE PRINTED AND PROCESSING STOPPED.

TITLE,EXAMPLE OF BACKWARD OPTION (DATA FROM DRAPER AND SMITH, P. 365-402) PAGE 4

SANDIA LABORATORIES <><> STEPWISE REGRESSION <><> FROM KANSAS STATE UNIVERSITY

AOV TABLE
ANALYSIS OF REGRESSION FOR VARIABLE 5---Y
(TABLE 1)

SOURCE	D.F.	SS	MS	F	SIGNIFICANCE
REGRESSION	4	2667.8994	666.97486	111.47917	.0000
RESIDUAL	8	47.863639	5.9829549		
TOTAL	12	2715.7631			

R**2 IS .98238
INTERCEPT IS 62.405369
STANDARD ERROR OF INTERCEPT IS 70.0710

VARIABLE NUMBER	VARIABLE NAME	REGRESSION COEFFICIENTS	STANDARDIZED REGRESSION COEFFICIENTS	PARTIAL SSQ	T-TEST VALUES 8 D.F.	R**2 DELETES	ALPHA HATS
1	X1	1.5511026	.606512	25.3503	2.0827	.9728	.1735
2	X2	.51016758	.527706	2.9725	.7049	.9813	.5013
3	X3	.10190940	.043390	.1091	.1350	.9823	.8960
4	X4	-.14406103	-.160287	.2470	-.2032	.9823	.8442

SINCE NO VARIABLES WERE FORCED TO STAY IN THE MODEL, THE VARIABLE WITH THE LARGEST $\hat{\alpha}$ > SIG. = .05 WILL BE DELETED.

UNIQUE SEQUENCE NUMBER FOR THIS ANOVA = 103

PRESS IS 110.

TITLE,EXAMPLE OF BACKWARD OPTION (DATA FROM DRAPER AND SMITH, P. 365-402) PAGE 5

SANDIA LABORATORIES <><> STEPWISE REGRESSION <><> FROM KANSAS STATE UNIVERSITY

AOV TABLE
ANALYSIS OF REGRESSION FOR VARIABLE 5---Y
(TABLE 1)

SOURCE	D.F.	SS	MS	F	SIGNIFICANCE
REGRESSION	3	2667.7903	889.25345	166.83168	.0000
RESIDUAL	9	47.972729	5.3303033		
TOTAL	12	2715.7631			

R**2 IS .98234
INTERCEPT IS 71.648307
STANDARD ERROR OF INTERCEPT IS 14.1424

VARIABLE NUMBER	VARIABLE NAME	REGRESSION COEFFICIENTS	STANDARDIZED REGRESSION COEFFICIENTS	PARTIAL SSQ	T-TEST VALUES } 9 D.F.	R**2 DELETES	ALPHA HATS
1	X1	1.4519380	.567737	820.9074	12.4100	.6801	.0001
2	X2	.41610976	.430414	26.7834	2.2418	.9725	.0340
4	X4	-.23654022	-.263183	9.9318	-1.3650	.9787	.2065

NEXT VARIABLE TO BE DELETED

UNIQUE SEQUENCE NUMBER FOR THIS ANOVA = 104

PRESS IS 85.4

TITLE,EXAMPLE OF BACKWARD OPTION (DATA FROM DRAPER AND SMITH, P. 365-402) PAGE 5

SANDIA LABORATORIES <><> STEPWISE REGRESSION <><> FROM KANSAS STATE UNIVERSITY

AOV TABLE
ANALYSIS OF REGRESSION FOR VARIABLE 5---Y
(TABLE 1)

SOURCE	D.F.	SS	MS	F	SIGNIFICANCE
REGRESSION	2	2657.8586	1328.9293	229.50370	.0000
RESIDUAL	10	57.904483	5.7904483		
TOTAL	12	2715.7631			

R**2 IS .97868
INTERCEPT IS 52.577349
STANDARD ERROR OF INTERCEPT IS 2.28617

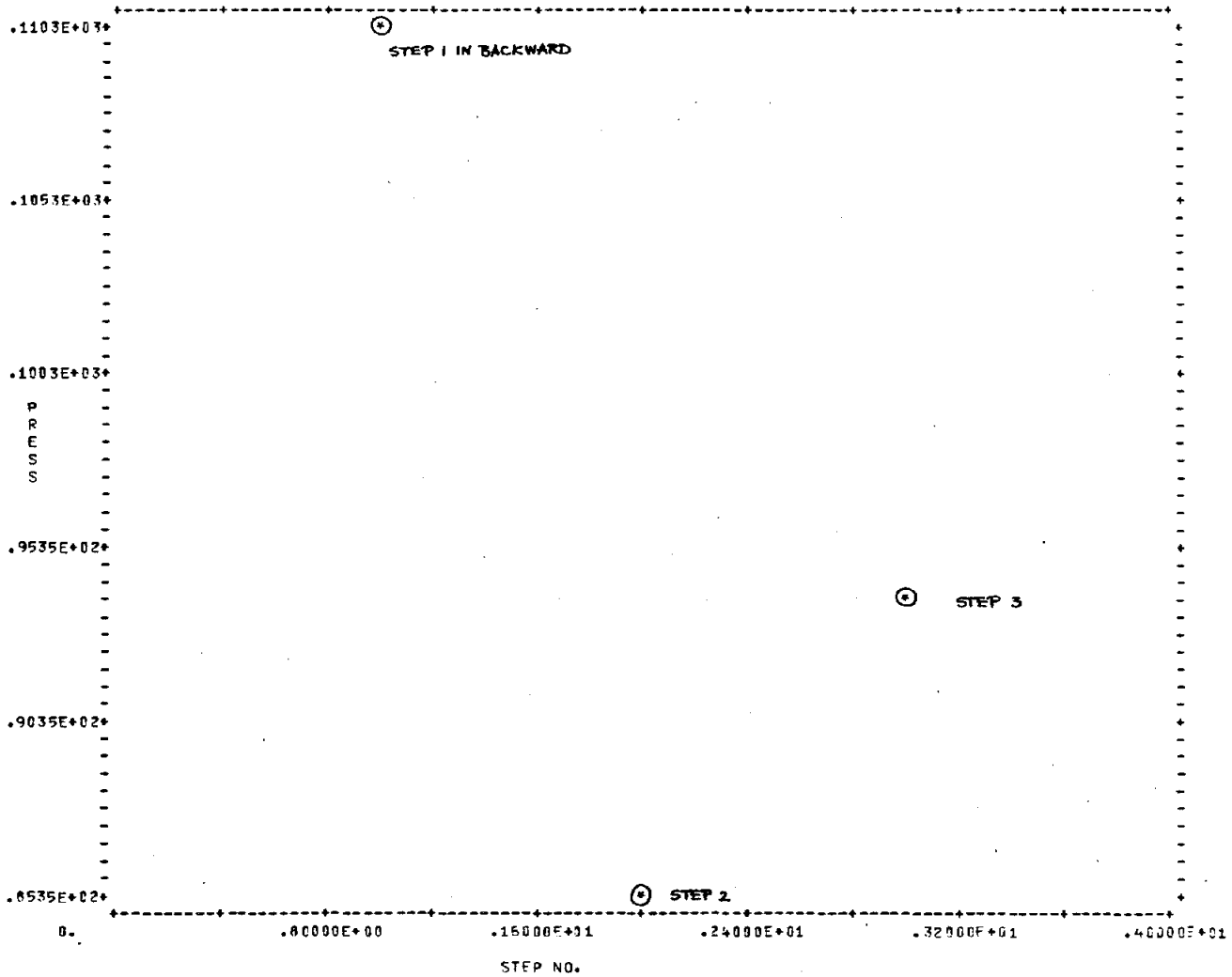
VARIABLE NUMBER	VARIABLE NAME	REGRESSION COEFFICIENTS	STANDARDIZED REGRESSION COEFFICIENTS	PARTIAL SSQ	T-TEST VALUES } 10 D.F.	R**2 DELETES	ALPHA HATS
1	X1	1.4683057	.574137	848.4319	12.1047	.6663	.0000
2	X2	.66225049	.685017	1207.7823	14.4424	.5339	.0000

BOTH < SIG = .05
⇒ NO MORE DELETIONS

UNIQUE SEQUENCE NUMBER FOR THIS ANOVA = 105

TITLE,EXAMPLE OF BACKWARD OPTION (DATA FROM DRAPER AND SMITH, P. 365-402) PAGE 7

SANDIA LABORATORIES <><> STEPWISE REGRESSION <><> FROM KANSAS STATE UNIVERSITY



TITLE,EXAMPLE OF BACKWARD OPTION (DATA FROM DRAPER AND SMITH, P. 365-402) PAGE 3

SANDIA LABORATORIES <><> STEPWISE REGRESSION <><> FROM KANSAS STATE UNIVERSITY

TABLE OF RESIDUALS FOR VARIABLES 5----Y

TIME	OBSERVED VALUE	PREDICTED VALUE	RESIDUAL
1	79.5000	80.8740	-1.57406
2	74.3000	73.2509	1.04908
3	104.300	105.815	-1.51474
4	87.6000	89.2585	-1.65848
5	95.9000	97.2925	-1.39251
6	109.200	105.152	4.04751
7	102.700	104.002	-1.30205
8	72.5000	74.5754	-2.07542
9	93.1000	91.2755	1.82451
10	115.900	114.538	1.36246
11	83.8000	80.5357	3.26433
12	113.300	112.437	.862756
13	109.400	112.293	-2.89344

END OF ANALYSIS FOR 2ND PASS ON THE DATA

SANDIA LABORATORIES <> STEPWISE REGRESSION PROGRAM <> COURTESY OF DEPT. OF STATISTICS - KANSAS STATE UNIVERSITY

TITLE,EXAMPLE OF RANK REGRESSION WITH THE STEPWISE OPTION(DATA FROM DRAPER/SMITH

DATA,5,0,2. (STAT CONTROL CARD)

INPUT CHECK OF PARAMETERS

NUMBER OF VARIABLES READ IN = 5

NO. OF TRANSFORMED VARIABLES = 0

DATA DISPOSITION IS 2

LABEL(1)=RANK(X1),RANK(X2),RANK(X3),RANK(X4),RANK(Y) → THESE LABELS WILL REPLACE THE LABELS USED ON THE PREVIOUS TWO PASSES. (STAT CONTROL CARD)

MODEL,5=1+2+3+4. (STAT CONTROL CARD)

STEPWISE,SIGIN=.05,SIGOUT=.10 (STAT CONTROL CARD)

PRESS (STAT CONTROL CARD)

RANK REGRESSION → A REGRESSION ANALYSIS WILL BE PERFORMED ON THE RANKS OF THE DATA. (STAT CONTROL CARD)

OUTPUT,CORR,STEPS,RESIDUALS. → SAME OUTPUT AS 2ND PASS. (STAT CONTROL CARD)

END OF PARAMETERS (STAT CONTROL CARD)

TITLE,EXAMPLE OF RANK REGRESSION WITH THE STEPWISE OPTION(DATA FROM DRAPER/SMITH PAGE 1

SANDIA LABORATORIES <><> STEPWISE REGRESSION <><> FROM KANSAS STATE UNIVERSITY

INPUT CHECK ON DATA

FIRST OBSERVATION

NOTE: THESE ARE THE RANKS ASSIGNED TO THE FIRST OBSERVATION. RANKS ARE ASSIGNED BY VARIABLE NUMBER.

RANK(X1)	RANK(X2)	RANK(X3)	RANK(X4)	RANK(Y)
1	2	3	4	5
6.5000000	1.0000000	2.5000000	13.000000	3.0000000

AVERAGE RANK RESULTING FROM TIED INPUT OBSERVATIONS.

NO. RAW DATA INPUT = 13

NO. TRANSFORMED OBSERVATIONS = 13

NO. OF OBSERVATIONS DROPPED = 0

TITLE,EXAMPLE OF RANK REGRESSION WITH THE STEPWISE OPTION(DATA FROM DRAPER/SMITH PAGE 2

SANDIA LABORATORIES <><> STEPWISE REGRESSION <><> FROM KANSAS STATE UNIVERSITY

VARIABLE NAME	VARIABLE NUMBER	MEAN	VARIANCE	STD. DEV.	STD. ERR.	C.V.
RANK(X1)	1	7.00000	14.5417	3.81335	1.05763	54.48
RANK(X2)	2	7.00000	15.1250	3.88909	1.07864	55.56
RANK(X3)	3	7.00000	14.9167	3.86221	1.07118	55.17
RANK(X4)	4	7.00000	15.0833	3.88373	1.07715	55.48
RANK(Y)	5	7.00000	15.1667	3.89444	1.08012	55.63

13 OBSERVATIONS

AVERAGE RANK ASSIGNED FOR 13 OBSERVATIONS IS 7.

WITHOUT TIED OBSERVATIONS THE VALUES IN THESE COLUMNS WILL ALL BE THE SAME WITHIN A COLUMN

TITLE,EXAMPLE OF RANK REGRESSION WITH THE STEPWISE OPTION(DATA FROM DRAPER/SMITH PAGE 3

SANDIA LABORATORIES <><> STEPWISE REGRESSION <><> FROM KANSAS STATE UNIVERSITY

CORRELATION MATRIX

RANK(X1)	1	1.0000				
RANK(X2)	2	.3301	1.0000			
RANK(X3)	3	-.7186	.0527	1.0000		
RANK(X4)	4	-.3320	-.9903	-.0806	1.0000	
RANK(Y)	5	.7912	.7373	-.4488	-.7521	1.0000
	NO.	1	2	3	4	5
NAME		RANK(X1)	RANK(X2)	RANK(X3)	RANK(X4)	RANK(Y)

RANK CORRELATIONS

TITLE.EXAMPLE OF RANK REGRESSION WITH THE STEPWISE OPTION(DATA FROM DRAPER/SMITH PAGE 4

SANDIA LABORATORIES <><> STEPWISE REGRESSION <><> FROM KANSAS STATE UNIVERSITY

AOV TABLE
ANALYSIS OF REGRESSION FOR VARIABLE 5---RANK(Y)
(TABLE 1)

SOURCE	D.F.	SS	MS	F	SIGNIFICANCE
REGRESSION	1	113.93123	113.93123	18.411433	.0013
RESIDUAL	11	68.068768	6.1880698		
TOTAL	12	182.00000			

R**2 IS .62600
INTERCEPT IS 1.3438395
STANDARD ERROR OF INTERCEPT IS 1.48783

VARIABLE NUMBER	VARIABLE NAME	REGRESSION COEFFICIENTS	STANDARDIZED REGRESSION COEFFICIENTS	PARTIAL SSQ	T-TEST VALUES	R**2 DELETES	ALPHA HATS
1	RANK(X1)	.80802292	.791199	113.9312	4.2909	0.0000	.0020

UNIQUE SEQUENCE NUMBER FOR THIS ANOVA = 106

RANK FIT GIVES A RAW DATA NORMALIZED R**2 = .64292846

COEFFICIENT OF INTERPOLATION = .68409359E-01

PRESS IS 90.347

THE USER SHOULD NOTE THAT THIS R**2 VALUE IS CALCULATED AS

$$\frac{\sum(\hat{y}_i-\bar{y})^2}{\sum(\hat{y}_i-\bar{y})^2+\sum(y_i-\hat{y}_i)^2}$$

AND SINCE THE $\hat{y}_i$ VALUES HAVE BEEN OBTAINED THROUGH INTERPOLATION THIS CALCULATION WILL NOT NECESSARILY RESULT IN THE SAME R**2 VALUE AS THE USUAL CALCULATION

$$\frac{\sum(\hat{y}_i-\bar{y})^2}{\sum(y_i-\bar{y})^2}$$

SINCE THE $\hat{y}_i$ VALUES ARE NOT THE LEAST SQUARES ESTIMATES, THE QUANTITIES $\sum(\hat{y}_i-\bar{y})^2$ AND $\sum(y_i-\hat{y}_i)^2$ ARE NOT ORTHOGONAL. THAT IS, THE SUM OF CROSS-PRODUCTS $\sum(y_i-\hat{y}_i)(\hat{y}_i-\bar{y})$ IS NON-ZERO. IF THE SUM OF CROSS-PRODUCTS IS NEAR ZERO, THEN THE COEFFICIENT OF INTERPOLATION WILL BE NEAR ZERO. IF IT IS LARGE, THEN THE COEFFICIENT WILL BE NEAR ONE. SEE THE TEXT FOR FURTHER DISCUSSION.

TITLE,EXAMPLE OF RANK REGRESSION WITH THE STEPWISE OPTION(DATA FROM DRAPER/SMITH PAGE 5

SANDIA LABORATORIES <><> STEPWISE REGRESSION <><> FROM KANSAS STATE UNIVERSITY

AOV TABLE
ANALYSIS OF REGRESSION FOR VARIABLE 3---RANK(Y)
(TABLE 1)

SOURCE	D.F.	SS	MS	F	SIGNIFICANCE
REGRESSION	2	162.92283	81.461413	42.700984	.0000
RESIDUAL	10	19.077175	1.9077175		
TOTAL	12	182.00000			

R**2 IS .89518
INTERCEPT IS 6.5099988
STANDARD ERROR OF INTERCEPT IS 1.31214

VARIABLE NUMBER	VARIABLE NAME	REGRESSION COEFFICIENTS	STANDARDIZED REGRESSION COEFFICIENTS	PARTIAL SSQ	T-TEST VALUES	R**2 DELETES	ALPHA HATS
1	RANK(X1)	.62154180	.608601	59.9822	5.6073	.5656	.0007
4	RANK(X4)	-.55154162	-.550024	48.9916	-5.0676	.6260	.0011

UNIQUE SEQUENCE NUMBER FOR THIS ANOVA = 107

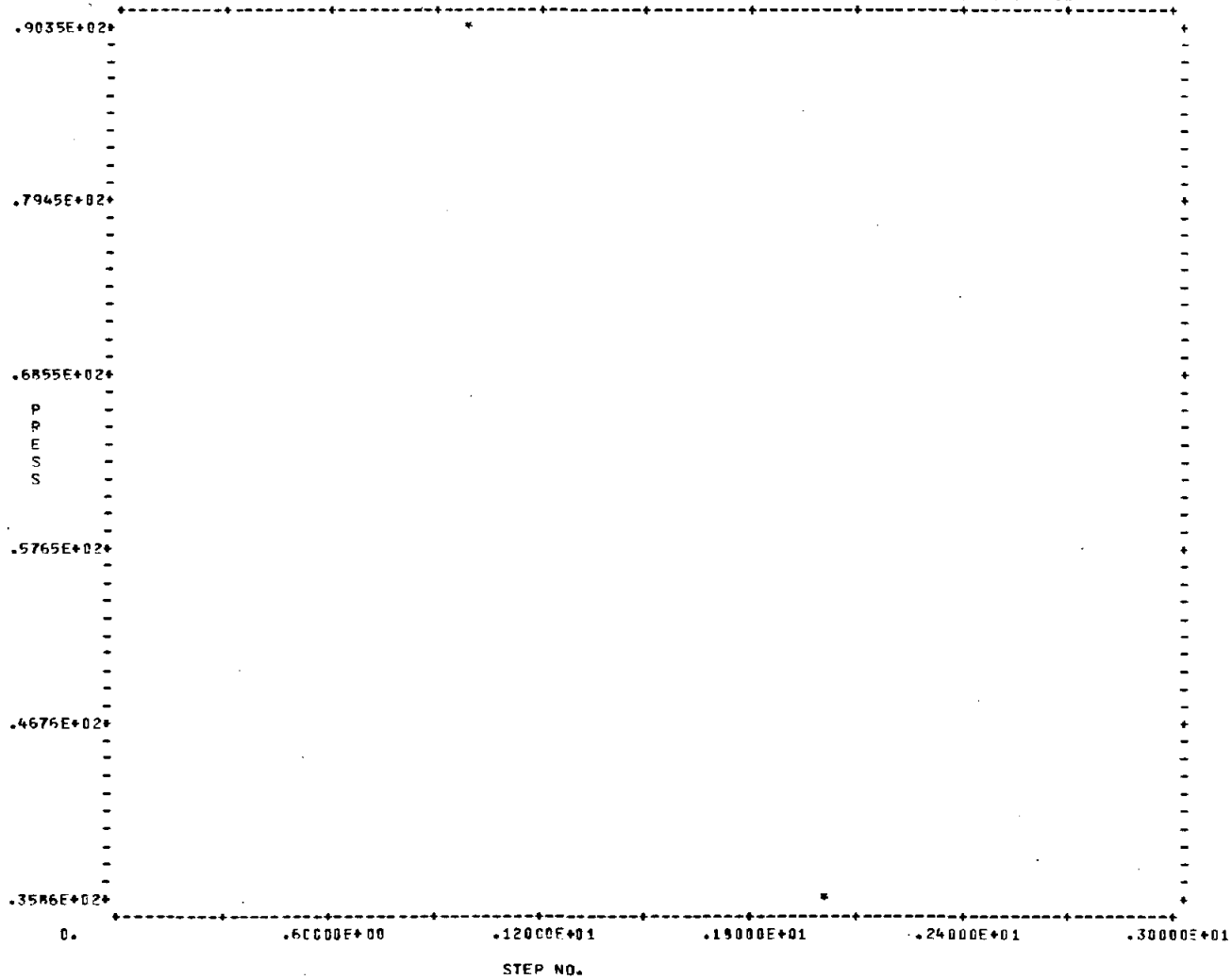
RANK FIT GIVES A RAW DATA NORMALIZED R**2 = .92844814

COEFFICIENT OF INTERPOLATION = .14538000E-01

PRESS IS 35.960

TITLE.EXAMPLE OF RANK REGRESSION WITH THE STEPWISE OPTION(DATA FROM DRAPER/SMITH PAGE 5

SANDIA LABORATORIES <><> STEPWISE REGRESSION <><> FROM KANSAS STATE UNIVERSITY



TITLE,EXAMPLE OF RANK REGRESSION WITH THE STEPWISE OPTION(DATA FROM DRAPER/SMITH PAGE 7

TABLE OF RESIDUALS FOR VARIABLES 5----RANK(Y)

TIME	RANK OF Y	PREDICTED RANK OF Y	RANK RESIDUAL	RAW Y	RAW YHAT	RAW RESIDUAL
1	3.0	3.37998	-.379979	78.5000	80.5139	-2.01393
2	2.0	1.13458	.865417	74.3000	72.7422	1.55775
3	9.0	10.8300	-1.83002	104.300	109.366	-5.06600
4	5.0	6.96923	-1.96923	87.6000	95.8138	-8.21384
5	7.0	6.13769	.862313	95.9000	93.4855	2.41448
6	10.0	10.0027	-.270874E-02	109.200	109.201	-.541747E-03
7	8.0	9.06617	-1.06617	102.700	104.624	-1.92421
8	1.0	2.23767	-1.23767	72.5000	75.2982	-2.79820
9	6.0	5.96269	.373130E-01	93.1000	92.8948	.205221
10	13.0	10.7293	2.27075	115.900	109.346	6.55415
11	4.0	2.78921	1.21079	83.8000	77.6147	6.18533
12	12.0	11.6573	.342666	113.300	111.964	1.33640
13	11.0	10.1035	.896521	109.400	109.221	.179304

THESE TWO COLUMNS CAN BE SAVED ON DISK 20

RESIDUAL SUM OF SQUARES ON RAW DATA = 200.052

TERMINATION ON END OF FILE - UNIT 5

SEE IMAN AND CONOVER (1979) FOR AN EXPLANATION OF HOW THESE RESIDUALS ARE CALCULATED.

```
SANDIA LABORATORIES <> STEPWISE REGRESSION PROGRAM <> COURTESY OF DEPT. OF STATISTICS - KANSAS STATE UNIVERSITY
TITLE-WEIGHT,TRANSFORMATION, AND DROP OPTIONS EXAMPLE - (DRAPER AND SMITH DATA)
DATA,6,3,1.                                                          (STAT CONTROL CARD)

                    INPUT CHECK OF PARAMETERS
                    NUMBER OF VARIABLES READ IN =   6
                    NO. OF TRANSFORMED VARIABLES =   3
                    DATA DISPOSITION IS  1
                 INPUT
                 VARIABLES
LABEL(1)=X1,X2,X3,X4,Y,WEIGHT                                        (STAT CONTROL CARD)
STEPWISE,SIGIN=.05,SIGOUT=.10                                        (STAT CONTROL CARD)
MODEL,5=1+2+3+4+7+8+9,                                               (STAT CONTROL CARD)
OUTPUT,STEPS                                                         (STAT CONTROL CARD)
DROP 8.                                                              (STAT CONTROL CARD)
TRANSFORMATION(7=1*1,8=4*4,9=1*4.                                    (STAT CONTROL CARD)
WEIGHT=6.                                                            (STAT CONTROL CARD)
END OF PARAMETERS                                                    (STAT CONTROL CARD)
FORMAT(6F6.0)
```

SAME 13 POINT DATA AS USED ON PREVIOUS STEPWISE EXAMPLE.

THERE ARE NO LABELS SPECIFIED FOR VARIABLES 7, 8, AND 9 AS THESE LABELS WILL BE ENCODED FROM THE TRANSFORMATION CARD.

WEIGHT IS JUST A CONVENIENT LABEL TO USE HERE AS THE WEIGHTS ARE ENTERED AS THE 6TH VARIABLE.

NOTE: VARIABLE NUMBER 6 DOES NOT APPEAR ON THE MODEL CARD

OBSERVATION NUMBER 8 WILL BE DROPPED

SPECIFIES WHICH OF THE INPUT VARIABLES IS TO BE USED AS THE WEIGHTS.

THIS CARD CREATES 3 NEW VARIABLES 7, 8, AND 9 WHICH ARE RESPECTIVELY EQUAL TO X_1^2, X_4^2, AND X_1X_4.

THIS EXAMPLE USES WEIGHTED REGRESSION. THE WEIGHTS MUST BE USER SUPPLIED FOR EACH OBSERVATION AND APPEAR AS ONE OF THE INPUT VARIABLES; IN THIS CASE THE 6TH VARIABLE POSITION CONTAINS THE WEIGHTS. FOR PURPOSES OF ILLUSTRATION THE RESIDUALS FROM THE PREVIOUS STEPWISE ANALYSIS ON THESE 13 DATA POINTS WERE SCALED SO THAT THE LARGEST RESIDUAL INABSOLUTE VALUE (5.02338) WAS REPRESENTED AS ZERO AND THE SMALLEST ABSOLUTE RESIDUAL (.137083) WAS REPRESENTED BY ONE. LINEAR INTERPOLATION WAS USED TO SCALE THE OTHER RESIDUALS. THESE SCALED VALUES WERE THEN USED AS WEIGHTS. PLEASE NOTE THAT THIS IS AN ENTIRELY ARBITRARY WAY TO ASSIGN WEIGHTS.

```
TITLE-WEIGHT,TRANSFORMATOON, AND DROP OPTIONS EXAMPLE - (DRAPER AND SMITH DATA)          PAGE  1
SANDIA LABORATORIES <><> STEPWISE  REGRESSION  <><> FROM KANSAS STATE UNIVERSITY
INPUT CHECK ON DATA
```

THESE LABELS ARE CREATED BY THE PROGRAM

FIRST OBSERVATION

X1	X2	X3	X4	Y	WEIGHT	1*1	4*4
1	2	3	4	5	6	7	8
7.0000000	26.000000	6.0000000	60.000000	78.500000	.59000000	49.000000	3600.0000

1*4
9
420.00000

$$1 - \left(\frac{2.16013 - .137083}{5.02338 - .137083} \right)$$

```
NO. RAW DATA INPUT =    13
NO. TRANSFORMED OBSERVATIONS =    12
NO. OF OBSERVATIONS DROPPED =    1 ---> OBSERVATION NUMBER 8 WAS DROPPED
WEIGHTED REGRESSION REQUESTED, WEIGHTS APPEAR AS VARIABLE NUMBER    6
```

TITLE-WEIGHT,TRANSFORMATCON, AND DROP OPTIONS EXAMPLE - (DRAPER AND SMITH DATA) PAGE 2

SANDIA LABORATORIES <><> STEPWISE REGRESSION <><> FROM KANSAS STATE UNIVERSITY

VARIABLE NAME	VARIABLE NUMBER	MEAN	VARIANCE	STD. DEV.	STD. ERR.	C.V.
X1	1	7.27406	37.8115	6.14911	1.77509	83.39
X2	2	50.2880	244.024	15.6213	4.50947	31.06
X3	3	12.0062	41.1702	6.41640	1.85226	53.44
X4	4	27.6908	286.038	16.9126	4.88226	61.08
1*1	7	85.0374	16184.2	127.217	36.7244	142.88
4*4	8	1028.98	.125483E+07	1120.19	323.372	108.86
1*4	9	188.768	36524.8	191.115	55.1701	101.24
Y	5	96.8853	157.960	14.0698	4.06161	14.52

12 OBSERVATIONS

THESE STATISTICS ARE ALL THE RESULT OF WEIGHTED CALCULATIONS, I.E.,

$$\text{WEIGHTED MEAN} = \frac{\sum w_i x_i}{\sum w_i}$$

COMPARISONS CAN BE MADE WITH THE PREVIOUS UNWEIGHTED EXAMPLE.

```
UNIQUE SEQUENCE NO. =    101        ANALYSIS FOR DEPENDENT VARIABLE  5---Y
TITLE,EXAMPLE OF STEPWISE OPTION (DATA FROM DRAPER AND SMITH, P. 365-402)
   1  1    0  .11756793E+03  4 -.73816181E+00

UNIQUE SEQUENCE NO. =    102        ANALYSIS FOR DEPENDENT VARIABLE  5---Y
TITLE,EXAMPLE OF STEPWISE OPTION (DATA FROM DRAPER AND SMITH, P. 365-402)
   2  1    0  .10309736E+03  1  .14399583E+01  4 -.61395363E+00

UNIQUE SEQUENCE NO. =    103        ANALYSIS FOR DEPENDENT VARIABLE  5---Y
TITLE,EXAMPLE OF BACKWARD OPTION (DATA FROM DRAPER AND SMITH, P. 365-402)
   4  1    0  .62405369E+02  1  .15511026E+01  2  .51016758E+00  3  .10190940E+00
   4  2    4 -.14406103E+00

UNIQUE SEQUENCE NO. =    104        ANALYSIS FOR DEPENDENT VARIABLE  5---Y
TITLE,EXAMPLE OF BACKWARD OPTION (DATA FROM DRAPER AND SMITH, P. 365-402)
   3  1    0  .71648307E+02  1  .14519380E+01  2  .41610976E+00  4 -.23654022E+00

UNIQUE SEQUENCE NO. =    105        ANALYSIS FOR DEPENDENT VARIABLE  5---Y
TITLE,EXAMPLE OF BACKWARD OPTION (DATA FROM DRAPER AND SMITH, P. 365-402)
   2  1    0  .52577349E+02  1  .14683057E+01  2  .66225043E+00

UNIQUE SEQUENCE NO. =    106        ANALYSIS FOR DEPENDENT VARIABLE  5---RANK(Y)
TITLE,EXAMPLE OF RANK REGRESSION WITH THE STEPWISE OPTION(DATA FROM DRAPER/SMITH
   1  1    0  .13438395E+01  1  .80802292E+00

UNIQUE SEQUENCE NO. =    107        ANALYSIS FOR DEPENDENT VARIABLE  5---RANK(Y)
TITLE,EXAMPLE OF RANK REGRESSION WITH THE STEPWISE OPTION(DATA FROM DRAPER/SMITH
   2  1    0  .65099988E+01  1  .62154180E+00  4 -.55154162E+00
```

EXAMPLE OF THE INFORMATION WRITTEN ONTO TAPE19.

TWO BLANK CARD IMAGES

CORRESPONDS TO UNIQUE SEQUENCE NO IN ANOVA TABLE

THIS IS THE SECOND CARD CONTAINING COEFFICIENTS

THE DEPENDENT VARIABLE IS VARIABLE NUMBER 5.

THE CONSTANT TERM AND ITS NUMBER "0".

TWO INDEPENDENT VARIABLES

THE FITTED REGRESSION EQUATION (EXCLUDING THE CONSTANT).

DISTRIBUTION:

W. J. Conover
College of Business Administration
Texas Tech University
Lubbock, Texas 79409

J. M. Davenport (5)
Department of Mathematics
Texas Tech University
Lubbock, Texas 79409

K. E. Kemp
Department of Statistics
Kansas State University
Manhattan, Kansas 66502

M. C. Cullingford (25)
NRC-PAS
Washington, D. C. 20555

1221	D. L. Wright
1223	R. R. Prairie
1223	H. E. Anderson
1223	C. R. Clark
1223	R. G. Easterling
1223	E. L. Frost (10)
1223	I. J. Hall
1223	R. L. Iman (25)
1223	D. D. Sheldon
1223	B. P. Roudabush
1223	M. Shortencarier (10)
1417	F. W. Muller
1417	E. E. Ard
1417	J. R. Ellefson
1417	F. W. Spencer
1763	M. D. Fimple
4413	J. E. Campbell
4413	R. M. Cranwell
4413	J. C. Helton
4723	K. A. Bergeron
4723	L. L. Lukens
4723	R. R. Peters
5641	G. P. Steck
8324	M. A. Griesel
8346	C. J. Decarli
9515	J. T. Hillman
9625	J. R. Yoder
3141	T. L. Werner (5)
3151	W. L. Garner (3) for DOE/TIC (Unlimited Release)
3154-3	R. P. Campbell (25) for DOE/TIC