REGRESS: A Biased Regression Package

by

Robert F. Pierce

Technical Report No. 124
Department of Statistics AFOSR Contract

October 1, 1976

Research sponsored by the Air Force Office
of Scientific Research
Contract 75-2871

Department of Statistics
Southern Methodist University
Dallas, Texas 75275

REGRESS

1. Scope of REGRESS

REGRESS is a linear multiple regression package consisting of programs capable of providing the computations necessary for Ordinary Least Squares Regression Analysis (OLS), Latent Root Regression Analysis (LRRA), Principal Component Regression Analysis (PCR), Ridge Regression Analysis (RRA), and a Baranchik (James-Stein) Shrunken Estimator (BSE). The user may call any or all of these analyses.

2. Purpose of REGRESS

The purpose of REGRESS is to bring OLS and four methods of biased regression together in one user-oriented package so that the user can more easily analyze data that is possibly subject to multicollinearities.

3. Background

3.1 Unbiased Regression

There are many multiple regression packages available [e.g.1,2,11]. These perform OLS and, in some cases, Weighted Least Squares to analyze the assumed relationship between a dependent or response variable (usually denoted Y) and a set of independent or regressor variables (X's). A modeling of this relationship is given by

$$Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \ldots + \beta_p X_{ip} + \varepsilon_i, \quad i = 1, \ldots, n \quad (1)$$

where Y_i is the <u>ith</u> observation of the response variable, Y;

X_{ij} is the <u>ith</u> value of the <u>jth</u> independent variable;

β_j is the <u>jth</u> regression coefficient;

ε_i is an unobservable error term.

Each ε_i is independently normally distributed, $E[\varepsilon_i] = 0$, $Var[\varepsilon_i] = \sigma^2$.

The goal is to estimate the β_j according to some criteria to yield a prediction equation:

$$\hat{Y} = b_o + b_1 X_1 + \ldots + b_p X_p.$$

Usually this estimation of the β_j is done using certain assumptions about (1) and a "least squares" criterion, the OLS procedure, which gives b_j that are <u>unbiased</u> estimators of the β_j. For more on the OLS approach to the regression problem, see [3].

3.2 Biased Regression

When the X_j are strongly "correlated", we say we have multicollinear data [9,13]. It can be shown that multicollinearities among the regression variables increases the variance of the estimators of the β_j coefficients, the b_j [e.g., see 9]. So while using OLS we would have unbiased estimators of the β_j, the variance of those estimators could be so large as to render the estimates of little use. This is the problem addressed by biased regression.

Now, the rationale of biased regression can be given in 2 steps:

(i) partition the estimator into recognizable components whose corresponding variances can be determined, and

(ii) reduce or remove the components, with larger variances, thereby reducing the total variance.

Four different ways of performing steps (i) and (ii) yield LRRA, PCR, RRA and BSE. All four of the methods yield estimators with reduced variance. As a consequence they also yield estimators that are biased as shown by the Gauss-Markov theorem [3,p59]. Only the user can decide whether the loss (bias) is worth the gain (reduced variance).

The uninformed user should acquaint himself with the underpinnings of these methods before using REGRESS. This is important not only because REGRESS calls upon the user to supply certain information needed in computation, but also because REGRESS itself cannot address the problem of which method of biased regression, if any, suits the user's problem.

An excellent overview of biased regression with a comprehensive bibliography is given in [5]. Also, individual references for each method are given later in the output description, Section 8.

4. The Control of REGRESS

4.1 REGRESS and SPSS

One of the main objectives of any utility package is ease-of-use or high user orientation. The success of the Statistical Package for the Social Sciences (SPSS) [11] in accomplishing this objective has greatly influenced the design of REGRESS to the extent that REGRESS control cards have been given the same general format as those of SPSS and the output format is somewhat like that of SPSS. This instruction sheet also was patterned after that of SPSS. For all those similarities there still are some major differences, however, because REGRESS and SPSS were written for different purposes. So the user familiar with SPSS, though possessing a distinct advantage, should still study the format of each REGRESS control card before trying a REGRESS run.

4.2 REGRESS Control Cards

All REGRESS control cards have two portions: (1) a control field which occupies card columns 1 to 15 and contains the control word(s) identifying the card to both the user and the package; and (2) the specification field which occupies card columns 16 to 80 of that and all subsequent cards necessary to complete the specification, and containing the parameters and arguments required by the particular control card used.

4.2.1 Control Field

Each card is identified by a unique control word or set of control words. These control words inform the computer as to the type of specification that will follow so that the proper procedures can be called up to act upon the information on the card.

The control word(s) always begin in column 1 and extend to column 15, if necessary. Spelling and spacing of these control words are crucial for obvious reasons.

4.2.2. Specification Field

Columns 16 to 80 are filled with information that conforms to the control field. Some of the control cards have free-field specification input, others are fixed-field. This is an important departure from SPSS control.

4.2.3 REGRESS Control Cards

```
                1                                                          8
1 Control Field6 ←————————— Specification Field ————————————————→          0
```

RUN NAME
N OF CASES
N OF VARS
Y-VAR
INPUT FORMAT
INPUT OPTION
METHODS
RESIDUALS
DLET VARS
VCTR DLET MAX L
VCTR DLET MAX P
VCTR DLET ORD L
VCTR DLET ORD P
RIDGE K-START
RIDGE K-STOP
RIDGE RESID K
READ INPUT DATA

The above are the 17 possible control cards. A typical run will include, in this order:

(1) computer system cards
(2) a subset of the REGRESS control cards shown above
(3) data cards
(4) an EOF card.

REGRESS cannot treat more than one set of these control cards on one run.

4.2.4 Similarities between REGRESS and SPSS Control Cards

(1) Control field is columns 1 to 15
(2) Specification field is columns 16 to 80
(3) Order-dependencies exist (Section 7)
(4) Continuation to another card is accomplished by leaving columns 1 to 15 blank and continuing the information in column 16 (see (3) below).

4.2.5 Differences between REGRESS and SPSS Control Cards

(1) REGRESS control words must be spelled out completely
(2) REGRESS has some fixed formats in the specification field
(3) Only REGRESS control cards that have free-field formatting in the specification field can be continued to a second card.
(4) REGRESS uses no delimiters.

4.2.6 Control Card Errors

4.2.6.1 Control Field Errors

Anytime REGRESS encounters a control word or words in columns 1 to 16 that is unrecognizable, an error message is printed which consists of the unrecognized card and a small diagnostic message. REGRESS then aborts. No endeavor is made to read subsequent cards since these diagnostics are many times misleading.

4.2.6.2 Specification Field Errors

Numerous error tests are performed on each control card. Formatting is given careful scrutiny since faulty left- or right-justification can lead to completely erroneous results.

Upon encountering either faulty formatting, omission, or bad parametric values, REGRESS prints an error message consisting of

(1) the faulty card and
(2) a diagnostic message.

REGRESS then aborts.

All common mis-formatting errors are tested for. There are about 50 different error situations that REGRESS can detect. REGRESS cannot, of course, test for all errors in parametric values; for this reason the user must excercise caution since the results of a run with bad parametric values are unpredictable.

5. Data Preparation and Input

Generally, data entered into the REGRESS package consists of <u>cases</u> or observations (see exception below). A case consists of a response value of the dependent variable with the values of the independent variables associated with that response. Let n be the number of cases to be entered and p the number of regressor variables. The limitations on n and p are:

(1) $2 \leq n \leq 150$

(2) $1 \leq p \leq 30$ and $p < n$

So, a case will consist of p + 1 numbers (1 for the dependent variable and p for the regressor variables). There will be n such cases. This does not imply that the user will only have n data cards. A case can be on as many contiguous cards as desired. There is, however, no provision to put more than one case per card. So there will always be

at least n cards in the data sequence (see exception below).
Formatting of the data cards will be discussed in Section 6.5.

Exception: REGRESS also allows data to be input from the correlation matrix. See INPUT OPTION card, Section 6.6.

6. Descriptions of REGRESS Control Cards

This section fully describes the formatting and parametric limitations of each REGRESS control card, its status (optional or mandatory) and its specification format (which pertains only to columns 16 to 80 of that card). Deck placement information can be found in Section 7.

6.1 RUN NAME card

The RUN NAME card identifies the current run. This user-supplied label is printed at the top of the second page of output. There may be as many as four continuation (five total) cards. Therefore, the label may be as long as 5x65 = 325 characters. Each continuation card is printed on a new line.

STATUS: Optional
SPECIFICATION FORMAT:
Free-field
LIMITATION:
Five or less cards

6.2 N OF CASES card

This card informs the system of the number of cases in the user's data. If the number of cases is unknown, the user may write UNKNOWN in columns 16 to 22 and REGRESS will count the cases. This UNKNOWN option is not available with the INPUT OPTION.

STATUS: Mandatory
SPECIFICATION FORMAT:
Right-justified in columns 16 to 18.
LIMITATIONS:
(1) 2 $\leq$ N OF CASES $\leq$ 150
(2) N OF CASES > Number of regressors

6.3 N OF VARS card

The N OF VARS card tells REGRESS the number of variables (both response and regressors) in each case. If a DLET VARS card is being used, this N OF VARS number includes all variables before deletion.

STATUS: Mandatory
SPECIFICATION FORMAT:
Right-justified in columns 16 and 17
LIMITATIONS:
(1) N OF VARS $\leq$ 31
(2) The number of regressors left after the DLET VARS (if any) must be less than N OF CASES.

6.4 Y-VAR card

At the start of a run, REGRESS is prepared to treat any of the variables on a case as the dependent variable. The Y-VAR card informs REGRESS which of the variables, ordinally numbered left to right from 1, in the case is the dependent variable. This placement of the dependent variable must remain constant throughout all n cases. So, if the dependent variable is listed second in the first case, it must be second in all succeeding cases. The Y-VAR card would then have a 2 in column 17.

STATUS: Mandatory

SPECIFICATION FORMAT:

Right-justified in columns 16 and 17.

LIMITATION:

Y-VAR $\leq$ N OF VARS

CAUTION:

DLET VARS should <u>not</u> be taken into consideration in determining the Y-VAR number. If Y-VAR is the thirteenth number in a case, Y-VAR is 13 in columns 16 and 17 whether or not any of the preceding twelve independent variables are deleted.

6.5 INPUT FORMAT card

The INPUT FORMAT card informs the system how to read a case. The information conveyed is exactly the same as that conveyed in a FORTRAN FORMAT statement. Free-field reading is also available by writing FREEFIELD in columns 16 to 24.

Those not familiar with FORTRAN FORMAT specifications may look to [10] or any other standard FORTRAN manual for assistance.

As in the RUN NAME card there can be as many as four continuation cards.

FREEFIELD is by far the easiest method to use and this fact should be kept in mind during data preparation.

If the INPUT OPTION is used the INPUT FORMAT must be FREEFIELD.

STATUS: Mandatory

SPECIFICATION FORMAT:

(1) If free-field input, FREEFIELD in columns 16 to 24.
(2) If fixed-field input, the first non-blank character in the field must be an open parenthesis. Following the parenthesis are regular FORTRAN FORMAT specifications. Imbedded blanks are OK. The last non-blank character must be a close parenthesis.

LIMITATION:

Five or less cards.

CAUTION:

(1) This formatting information pertains only to a single case. All other cases, of course, must be read in the same format.
(2) Unlike SPSS, the word "FIXED" is <u>not</u> used in REGRESS.

6.6 INPUT OPTION card

Many times data comes to a researcher in correlation form. The actual cases are either not handy or not available. REGRESS will take data input in this form with some loss of output power. The data must be input in the following order:

(1) Either the upper or lower triangle, ignoring the diagonal elements, of the "correlation" matrix (standardized X'X matrix). These elements are input in row sequence (left to right). So, if $X'X = M = [m_{ij}]$ and the input was the lower triangle, the order sequence would be

$m_{21}, m_{31}, m_{32}, m_{41}, m_{42}, m_{43}, \ldots, m_{p,p-1}$.

For the upper triangle the order would be

$m_{12}, m_{13}, m_{14}, \ldots, m_{1p}, m_{23}, m_{24}, \ldots, m_{2p}, \ldots, m_{p-1,p}$.

Note that the diagonal elements (always 1's) are not entered.

(2) The standardized $X'\underline{Y}$ vector. Note that both X and $\underline{Y}$ must be standardized. The order of the elements must, of course, be consistent with the information supplied about the correlation matrix.

(3) The vector of means of the independent variables.

(4) The mean of the dependent variable.

(5) The vector of square roots of the sum of squared deviations from the mean of each independent variable. The jth term in this vector is

$$\left[\sum_{i=1}^{n}(X_{ij} - \bar{X}_j)^2\right]^{1/2}$$

(6) The square root of the sum of squared deviations from the mean of the dependent variable. This is

$$\left[\sum_{i=1}^{n}(Y_i - \bar{Y})^2\right]^{1/2}$$

The number of values required for each step above is:

(1) $(\frac{1}{2})(p^2-p)$
(2) p
(3) p
(4) 1
(5) p
(6) 1

Yielding a total of $(\frac{1}{2})(p^2 + 5p + 4)$ values. These values are entered using free-field formatting and therefore do not have to conform to any input rules other than the ordering of the values. In particular, it is not necessary to start a new card when beginning each of the 6 steps.

CAUTION:

Careful attention to formatting of this card is especially important. Blanks are read as zeroes, so failure to right-justify can lead to unintended deletion of variables. REGRESS error-testing of this card is minimal.

6.11 RIDGE K-START and RIDGE K-STOP cards

REGRESS' Ridge Regression Analysis (RRA) uses a parameter k. This k changes and at each change a new analysis is done. The user can specify a closed interval, [a,b], in which k is confined by using the RIDGE K-START and RIDGE K-STOP cards.

RIDGE K-START specifies a in the interval, RIDGE K-STOP specifies b. (REGRESS partitions this interval into twenty (20) equal parts, yielding twenty-one (21) values for k.)

However, if the user only wants to see RRA for a single value of k, k_0, both RIDGE K-START and RIDGE K-STOP must specify this k_0. REGRESS, of course, would perform only one analysis at this k_0.

The default values of a and b are 0.00 and 0.20. These values are used when both cards are missing. If the user only includes one card, the undefined endpoint is taken to be the default value.

To determine the values of k the user has included by specifying the interval [a,b]:

k takes on values defined by

$$k_i = a + (i - 1)\left(\frac{b-a}{20}\right) \text{ for } i = 1,2,\ldots,21.$$

It is not necessary for one of these k_i to be equal to the k specified on a RIDGE RESID K card.

STATUS: Optional

SPECIFICATION FORMAT:

A floating-point or integer format number, right-justified in columns 16 to 25.

LIMITATION:

RIDGE K-START $\leq$ RIDGE K-STOP

6.12 VCTR DLET MAX L card

Latent Root Regression Analysis (LRRA) computes sets of regression coefficients after deleting a subset of latent vectors and latent roots of A'A[14]. Usually (i.e. without a VCTR DLET ORD L card) these subsets are (in order of deletion):

(1) The latent vector of A'A corresponding to the smallest latent root of A'A.
(2) The latent vector of A'A deleted in (1) and the latent vector corresponding to the second smallest latent root of A'A.
(3) The latent vectors of A'A deleted in (2) and the latent vector corresponding to the third smallest latent root of A'A.

etc.

Rarely would the user want to continue deletions until there are p latent vectors deleted. The user specifies on the VCTR DLET MAX L card the maximum number of latent vectors of A'A to be deleted.

The default value of VCTR DLET MAX L is 1.

STATUS: Optional
SPECIFICATION FORMAT:
Integer, right-justified in columns 16 and 17.
LIMITATION:
VCTR DLET MAX L < p

6.13 VCTR DLET ORD L card

As in 6.12, the usual order of deletion of latent vectors of A'A is determined by the size of the latent roots of A'A.

With the VCTR DLET ORD L card, however, the user can specify the deletion order by merely listing the desired order, using the following numerical naming scheme for the vectors:

0 for the latent vector of A'A associated with the smallest latent root of A'A
1 for the latent vector of A'A associated with the second smallest latent root of A'A
⋮
p for the latent vector of A'A associated with the largest latent root of A'A, where p is the number of regressors.

If this card is used, the ordering specified must satisfy the VCTR DLET MAX L value. In other words, if the user wishes to delete 3 latent vectors of A'A, the VCTR DLET MAX L card would have a 3 in column 17 (see Section 6.12). And now, if the user wants first to delete the latent vector associated with the second smallest latent root and, subsequent to that, to pick up with the regular deletion order, the VCTR DLET ORD L card would have a 1 in column 17, 0 in column 19, 2 in column 21. The point here is that the VCTR DLET ORD L card must contain three (3) numbers since the VCTR DLET MAX L was 3.
STATUS: Optional
SPECIFICATION FORMAT:
1st deleted vector is right-justified in columns 16 and 17;
2nd deleted vector is right-justified in columns 18 and 19;
3rd deleted vector is right-justified in columns 20 and 21;
⋮ ⋮
tth deleted vector is right-justified in columns 2t + 14 and 2t + 15 where t is the number specified on the VCTR DLET MAX L card.

6.14 VCTR DLET MAX P

Principal Components Regression Analysis (PCR) computes sets of regression coefficients after deleting a subset of latent vectors and latent roots of X'X[8]. Usually (i.e. without a VCTR DLET ORD P card) the subsets are (in order of deletion):

(1) The latent vector of X'X corresponding to the smallest latent root of X'X.
(2) The latent vector of X'X deleted in (1) and the latent vector or X'X corresponding to the second smallest latent root of X'X.
(3) The latent vectors of X'X deleted in (2) and the latent vector of X'X corresponding to the third smallest latent root of X'X.
etc.

Rarely would the user want to continue deletions until there are p-1 latent vectors deleted. The user specifies on the VCTR DLET MAX P card the maximun number of latent vectors of X'X to be deleted.

The default value of VCTR DLET MAX P is 1.

STATUS: Optional

SPECIFICATION FORMAT:

Integer, right-justified in columns 16 and 17.

LIMITATION:

VCTR DLET MAX P < p

6.15 VCTR DLET ORD P

As in 6.13, the usual order of deletion of latent vectors of X'X is determined by the size of the latent roots of X'X.

With the VCTR DLET ORD P card, however, the user can specify the order of deletion by merely listing the desired order, using the following numerical naming scheme for the vectors:

1 for the latent vector of X'X associated with the smallest latent root of X'X.

2 for the latent vector of X'X associated with the second smallest latent root of X'X.

⋮ ⋮

p for the latent vector of X'X associated with the largest latent root of X'X.

If the card is used, the ordering specified <u>must</u> satisfy the VCTR DLET MAX P value. In other words, if the user wishes to delete 3 latent vectors of X'X, the VCTR DLET MAX P card would have a 3 in column 17 (Section 6.14). And now, if the user wants first to delete the latent vector associated with the second smallest latent root and, subsequent to that, to pick up with regular deletion order, the VCTR DLET ORD P card would be used with a 2 in column 17, 1 in column 19, 3 in column 21. The point here is that the VCTR DLET ORD P card must contain three (3) numbers since the VCTR DLET MAX P card was 3.

STATUS: Optional

SPECIFICATION FORMAT:

1st deleted vector is right-justified in columns 16 and 17;
2nd deleted vector is right-justified in columns 18 and 19;
3rd deleted vector is right-justified in columns 20 and 21;

⋮ ⋮

rth deleted vector is right-justified in columns 2r + 14 and 2r + 15 where r is the number specified on the VCTR DLET MAX P card.

6.16 READ INPUT DATA card

The READ INPUT DATA card informs REGRESS that there are no more control cards to read and that the next cards to be encountered are data cards. This card also puts REGRESS into a final error-testing mode. Comments may be written in columns 16 to 80 without harm as this specification field is free. The comments will be printed along with the control field at the beginning of the REGRESS run. There is no provision to continue the comments on another card.

STATUS: Mandatory
SPECIFICATION FORMAT: none
LIMITATION:
This card must be the last control card.

7. Precedence Table

This section describes the order dependencies among REGRESS' control cards. No card may be placed after a card listed below it. Cards listed at the same precedence number may be ordered in any manner with respect to the cards at that number. The RUN NAME card may appear anywhere in the deck before the READ INPUT DATA card.

(1) N OF CASES
    N OF VARS
(2) Y-VAR
(3) METHODS
    RESIDUALS
    INPUT FORMAT
    INPUT OPTION
    VCTR DLET MAX L
    VCTR DLET MAX P
    DLET VARS
    RIDGE K-START
    RIDGE RESID K
(4) RIDGE K-STOP
    VCTR DLET ORD L
    VCTR DLET ORD P
(5) READ INPUT DATA

8. REGRESS output

8.1 Typical output

This section describes the output of a "full" REGRESS run (one that calls METHODS L, P, R and S and RESIDUALS for each method). The numbering below is for ordering and is not meant to designate pages.

(1) A listing of REGRESS control cards.
    This is where error messages would appear also.
(2) Run name as specified.
(3) Input data-response variable and independent variables by case. Not present if INPUT OPTION is used.
(4) Summary statistics for the input variables.
    (a) Means
    (b) Square root of the sum of squared deviations from the mean of each variable.
(5) Standardized X'X matrix (correlation form)
(6) Standardized X'Y vector (correlation form).
(7) Latent roots of X'X.
(8) Latent vectors of X'X

(9) Inverse of X'X

(10) Ordinary Least Squares Analysis [3]

- (a) Coefficients for prediction equation and their standard errors
- (b) ANOVA table
- (c) Statistics for adjusted regressors
- (d) Residuals, standardized residual plots and residual statistics. Not available with INPUT OPTION.

(11) Latent Root Regression Analysis [14]

- (a) Latent roots of A'A
- (b) Latent vectors of A'A
- (c) 1 latent vector deleted yields
  - (i) Coefficients for prediction equation
  - (ii) ANOVA table
  - (iii) A comparison of the standardized coefficients from OLS and LRRA
  - (iv) Statistics for adjusted regressors
  - (v) Residuals, standardized residual plots and residual statistics. Not available with INPUT OPTION.
- (d) 2 latent vectors deleted yields
  - (i) through (v) as in (c)

  ⋮

  LRRA continues in this manner until VCTR DLET MAX L is satisfied or SSR becomes negative (see Section 8.3).

(12) Principal Components Regression Analysis [8]

- (a) 1 latent vector deleted yields
  - (i) Coefficients for prediction equation and their standard errors.
  - (ii) ANOVA table
  - (iii) A comparison of the standardized coefficients from OLS and PCR
  - (iv) Statistics for adjusted regressors
  - (v) Residuals, standardized residual plots and residual statistics. Not available with INPUT OPTION.
- (b) 2 latent vectors deleted yields
  - (i) through (v) as in (a)

⋮

PCR continues in this manner until VCTR DLET MAX P is satisfied.

(13) Ridge Regression Analysis [6,7]

- (a) $k = a$ or k_0 (see Section 6.11) yields
  - (i) Coefficients for prediction equation and their standard errors.
  - (ii) ANOVA table
  - (iii) A table of comparisons and differences to aid in the choice of k.

(b) $k = a + \frac{(b - a)}{20}$ yields

(i) through (iii) as in (a)

⋮

RRA continues until k = b or, in the case of a = b, stops at step (a).

⋮

(v) k = RIDGE RESID K yields

(i) Residuals, standardized residual plots and residual statistics for k = RIDGE RESID K. Not available with INPUT OPTION.

(14) Baranchik (James-Stein) Shrunken Estimator [12]

(a) Coefficients for prediction equation

(b) ANOVA table

(c) Residuals, standardized residual plots and residual statistics. Not available with INPUT OPTION.

8.2 Computation of residual plots and statistics

Residuals, $e_i = Y_i - \hat{Y}_i$ are computed and listed when the RESIDUALS card is present. To plot these residuals, REGRESS "standardizes" them with the "unit normal deviate form" [3,p88]. The formula used to compute the ith standardized residual, e_i' is

$$e_i' = e_i/\hat{\sigma}$$

where $\hat{\sigma} = \sqrt{MSE}$. Now, an important point is that the MSE used by REGRESS for the above formula is the MSE found in the ANOVA table corresponding to the current computation of $\hat{Y}$. This yields a "pooled" MSE. Some argue that MSE should be the MSE from OLS regardless of the method used and the resulting SSR. The user must decide which method is preferable; making the obvious transformation if the decision is in favor of the unpooled MSE.

The residual plots are not used to compute the residual statistics as the plots are subject to the round-off necessary in plotting. For this reason, the number of runs, for example, may not be accurately portrayed in the plotting.

8.3 Termination of Latent Root Regression Analysis

REGRESS automatically terminates LRRA if the sum of squares due to regression becomes negative. Upon termination an explanatory message is printed.

8.4 Warning message for Ridge Regression Analysis

REGRESS provides a warning message if the sum of squares due to regression (SSR) becomes negative. Computation continues, however.

```
Y-VAR           13
RUN NAME        GORMAN-TOMAN DATA
                10/1/76
METHODS         LPRS
INPUT FORMAT    FREEFIELD
RIDGE K-START           .1
VCTR DLET MAX L 5
VCTR DLET MAX P 4
DLET VARS       1112
RESIDUALS       LPRS
RIDGE RESID K         2.0
RIDGE K-STOP          2.1
READ INPUT DATA
—cases—
6/7/8/9
```

12. Programming Information

REGRESS was written in the FORTRAN Extended language for Southern Methodist University's CDC CYBER-72. The system used was KRONOS 2.1, level 393.

REGRESS requires field lengths of 111700 to load and 104100 to run.

Compatibility with other systems cannot be measured, but a major consideration is REGRESS' use of the ENCODE and DECODE statements. These are not standard FORTRAN statements and their equivalents in standard FORTRAN would require careful programming.

REGRESS' main program directs sixteen subprograms. In brief, the main program reads and decodes the control cards. It then calls the subprograms according to the control card information.

The source program for REGRESS is 2232 80-character cards long.

REGRESS' subroutine EIGEN is from the System /360 Scientific Subroutine Package (360A - CM - 03X, New York: IBM Technical Publications Department, August 1970).

13. Acknowledgment

The development of REGRESS was sponsored in part by the Air Force Office of Scientific Research, Air Force Systems Command, USAF, under Grant No. AFOSR-75-2871.

References

[1] Dixon, W.J. (Ed.)(1973). BMD. University of California Press, Los Angeles.

[2] Dixon, W.J. (Ed.)(1975). BMDP. University of California Press, Los Angeles.

[3] Draper, N.R. and Smith, H.(1966). Applied Regression Analysis. Wiley, New York.

[4] Gorman, J.W. and Toman, R.J.(1966). Selection of Variables for Fitting Equations to Data, Technometrics, 8, 27-51.

[5] Hocking, R.R.(1976). The Analysis and Selection of Variables in Linear Regression, Biometrics, 32, 1-49.

[6] Hoerl, A.E. and Kennard, R.W. (1970). Ridge Regression: Applications to Non-Orthogonal Problems, Technometrics, 12, 69-82.

[7] Hoerl, A.E. and Kennard, R.W. (1970). Ridge Regression: Biased Estimation for Non-Orthogonal Problems, Technometrics, 12, 55-68.

[8] Mansfield, E.R. (1975). "Principal Component Approach to Handling Multicollinearity in Regression Analysis," Ph.D. dissertation, Southern Methodist University, Dallas.

[9] Mason, R.L., Webster, J.T., and Gunst, R.F. (1975). Sources of Multicollinerarity in Regression Analysis, Comm. in Statist., 4, 277-292.

[10] McCracken, D.D. (1972). A Guide to Fortran IV Programming, 2nd ed., Wiley, New York.

[11] Nie, N.H., Hull, C.H., Jenkins, J.G., Steinbrenner, K., Bent, D.H. (1975). Statistical Package for the Social Sciences, 2nd ed., McGraw-Hill, New York.

[12] Sclove, S.L. (1968). Improved Estimators for Coefficients in Linear Regression, Journal of the American Statist. Assoc., 63, 596-606.

[13] Silvey, S.P. (1969), Multicollinearity and Imprecise Estimation, Journal of the Royal Statist. Society B, 31, 539-552.

[14] Webster, J.T., Gunst, R.F., and Mason, R.L. (1974). Latent Root Regression Analysis, Technometrics, 16, 513-522.