

THE ECONOMIC
AND SOCIAL
RESEARCH INSTITUTE
MEMORANDUM SERIES
NO.76

REGRECON - A MULTIPLE-OPTION REGRESSION PROGRAM FOR
USE IN ECONOMETRIC RESEARCH

Peter Neary

Confidential: Not to be quoted
until the permission of the Author
and the Institute is obtained.

REGRECON - A MULTIPLE-OPTION REGRESSION PROGRAM FOR
USE IN ECONOMETRIC RESEARCH

1 Introduction

REGRECON is a card-based computer program for multiple regression analysis designed to provide a flexible tool for econometric research. Single equation ordinary least squares (O.L.S.) is the only estimating method for which explicit provision is made; however a range of options is provided, which permit the user to deal with a variety of problems commonly encountered in practical econometric applications.

The following sections of this memorandum describe the background and special features of the program. Readers who are solely interested in learning how to submit problems to the program should study the instructions in Section 4, as well as the examples of Section 5. Since the program is designed for use by practising econometricians, no knowledge of computer programming is required, though some familiarity with the preparation of punched card input would be an advantage.

2 Computational Considerations

REGRECON is a considerably revised and extended version of the REGRE program, described on pages 404-408 of IBM Manual "System / 360 SSP" [4]. The basic mathematical operation carried out by the program is the calculation of OLS estimates of the coefficient vector y in the model:

$$y = Z\gamma + u$$

given a sample of T observations on a dependent variable y and $(K+1)$ independent variables; γ and u are unobservable vectors of $(K+1)$ coefficients and T disturbance terms respectively. Z and y may be partitioned as follows:

$$Z = \begin{bmatrix} i & X \end{bmatrix} \quad y = \begin{bmatrix} -\alpha \\ \beta \end{bmatrix}$$

where i is a column-vector of ones, α is the intercept, and X is a $(T \times K)$ matrix of T observations on K independent variables.

Methods of estimating y , and the properties of such estimates are assumed to be familiar to the reader (for a review, see, e.g., Johnston [5], chap. 4). Suffice it to say that the normal method of calculating an estimate of y is to first estimate β using:-

$$\hat{\beta} = (X^1 X)^{-1} X^1 y$$

where all variables are now measured as deviations from their means. An estimate $\hat{\alpha}$ is then derived from $\hat{\beta}$. A refinement introduced in REGRE and retained in the REGRECON program is to first standardize all variables by multiplying them by $\sqrt{T-1}$ times their standard deviations. Letting S equal a $(K \times K)$ diagonal matrix whose elements are the standard deviations of the X 's multiplied by $\sqrt{T-1}$, and s_y equal the standard deviation of y multiplied by $\sqrt{T-1}$, write:

$$X^* = X S^{-1}, \quad y^* = y s_y^{-1}$$

The O.L.S. estimates can now be derived by first calculating the so-called "beta coefficients":

$$b = (X^{*1} X^*)^{-1} X^{*1} y^*$$

and then calculating:

$$\hat{\beta} = S^{-1} b s_y$$

The usual test statistics, R , F and t , are calculated in a similar manner (for details see [4] p.37).

The advantage of this roundabout procedure derives from the fact that the matrix which must be inverted to calculate b is in fact the correlation matrix of the X 's. Since the determinant of this matrix can only vary between zero and unity (irrespective of the units of measurement of the original

variables), this method greatly reduces the probability of severe rounding error. The results may be seen in a study by Longley [8], which compares the accuracy of a number of different regression programs, including the sub-routines used by REGRE. The latter, while not the most accurate of all the programs tested, were found to compare favourably with the others, and could be relied on to be accurate to at least four significant digits. (Despite these remarks, the user is warned that the probability of rounding error increases the closer the data matrix approaches singularity - i.e., the higher the intercorrelations between all variables. In such a situation, it is recommended that the variables be standardized before submitting the problem to the computer and the results checked by re-running the problem with a constant multiple of one of the independent variables added to the dependent variable, as suggested by Mullet and Murray [9]).

The only remaining computational problem encountered by the program is the calculation of a homogeneous regression equation (i.e. an equation where the intercept is constrained to equal zero). Instead of calculating the required coefficient vector $\hat{\beta}_h$ directly via the gross cross product matrix, the program again reduces the probability of rounding error by using an algorithm suggested by Stewart [10]. This first calculates the ordinary non-homogenous coefficient vector $\hat{\beta}$, and then derives $\hat{\beta}_h$ from the formula:

$$\hat{\beta}_h = \hat{\beta} + \frac{\hat{\alpha}}{\left\{ \frac{1}{T} + \bar{x}^{-1} (X^{*1} X^*)^{-1} \bar{x} \right\}} \{ (X^{*1} X^*)^{-1} \bar{x} \}$$

where \bar{x} is a $K \times 1$ vector of the means of the independent variables. Finally, the error sum of squares is calculated directly from the estimated residuals, and from it the usual test statistics and analysis of variance table are calculated in turn.

3. Features of the Program

Unless some of the special options available with REGRECON are requested by the user, the program operates in exactly the same way as the standard IBM program REGRE. These two programs are completely compatible, in the sense that data prepared for use with either one may be submitted to the other.

When no options are requested, the following information is printed out for each equation estimated:

1. Means, standard deviations, correlation with the dependent variable and estimated coefficient, its standard error and t-value, for each independent variable (including the intercept).
2. Multiple correlation coefficient R, adjusted R squared \bar{R}^2 , standard error of estimate, and a full analysis of variance table.
3. Test for multicollinearity: when the number of independent variables is greater than one, the determinant of the (standardised) X^1X matrix and a simple transformation of it are both printed out. The latter is a test statistic for the presence of singularity suggested by Haitovsky [3] and distributed as a chi-squared with degrees of freedom which are also printed out.
4. Tests for autocorrelation: The number of positive and negative residuals, and the values of Geary's tau (see [2]) and Durbin and Watson's d statistic are printed out.

In addition to this standard output, a number of different options may be specially requested:

- 1 Data Input: Data may be read in either variable by variable, or observation by observation. Once read in, it can be subjected to a wide range of transformations;

e.g., logarithms can be taken, lagged, first differenced, or moving average values can be generated, or variables can be seasonably corrected, using Leser's quasi-linear trend method (see [6] and [7]). For a full list of transformations available, see Appendix 2.

2. Data and Correlation Matrices for the full data set (including any transformed variables) may be printed out.

3. Table of Residuals may be printed out for each equation. In addition, for equations where the dependent variable is in log form, the antilogs of the actual and predicted y values may be printed out, as well as a new set of residuals calculated from them.

4. Re-estimation of Equation in Homogeneous Form: This may be done for each equation: a full print-out, including optional table of residuals is provided. Users should beware of applying this option when variables which are measured from arbitrary origins, e.g., time trends, are among the independent variables.

5. Summary Table:

This may be printed out at the end of each problem, listing the principal statistics associated with each equation as well as the independent variables included in each (though not their estimated coefficients). In addition the equations are ranked by the size of their adjusted R squared. If required the output referring to the original equations may be suppressed, and only the summary table for a particular problem printed.

6. Re-estimation of Equation with Different Set of Observations

In any one problem all the equations estimated refer to the same set of observations. However the same data can be used in the subsequent problem, this time with a different

set of observations. This may be repeated any number of times, without any new data being read in. An obvious application of this facility is the re-estimation of an equation over two sub-periods, with a view to testing for significant differences in the coefficients using the Chow test [5] pp.136-8; see section 5, example 2 below.

This completes the list of options available with the program. Note finally that tracing any errors in the operation of the program is facilitated by two devices: First of all, the parameters for each problem are printed out at the start, which should help in deciding whether the control card (see below) has been punched correctly. Secondly, any difficulty will usually be heralded by the printing out of one of the error messages given in Appendix 3.

4. How to Prepare Input to the Regression Program

First some jargon: for the purposes of this program, each time the program is used we call a run. Every run consists of one or more problems, each of which in turn normally requires the estimation of a number of equations or selections. There is no limit on the number of problems which can be included in each run and each problem normally consists of the following cards:

- A. Control Card (obligatory): This is the most important card of the problem, since it supplies the computer with the problem parameters. For details on how to prepare this card see Appendix 1. The values of all these parameters, both as read in and after subsequent adjustments, are printed out at the beginning of each problem. This should facilitate tracing the source of any error in the operation of the program.
- B. Transformation Cards (optional): These cards permit the user to request transformation on data read in for the problem, or in some cases to generate completely new data (e.g. time trends

or seasonal dummy variables). The number of transformation cards included must equal the value of the parameter 'MTRANS' on the control card.

In general the order of the transformation cards is irrelevant. There are exceptions to this rule however. Firstly, because of the structure of the program, transformations which require seasonal correction of input data (transformation 51) are carried out first, irrespective of the order in which the transformation cards occur. Secondly, note that transformations are cumulative: thus it is possible to read in two variables X1 and X2, add them together using transformation 2, to give (X1 + X2), and then take logs of the new composite variable, using transformation 7, to give log (X1 + X2). Obviously in this case the order of the transformation cards (which determines the order in which the operations are carried out) is crucial for the final result.

One transformation card must be included for each transformation required (the example of the last paragraph would count as two transformations; similarly reading in two variables and taking logs of both of them to give log X1 and log X2, would also count as two transformations); and each card must contain the following information in FORMAT (4I2,F6.0), i.e. four two column integers, followed by one six-column real variable:

<u>Columns</u>	<u>Variable</u>	<u>Meaning</u>
1-2	KCODE	= Transformation code (see Appendix 2)
3-4	MNEW	= Subscript of new (transformed) Variable
5-6	MOLD	= Subscript of pre-existing variable
7-8	M2	= Subscript of auxiliary pre-existing variable (only required for some transformation)
9-14	R	= Constant (only required for some transformations)

(If R is left blank, the following values are assumed:

0 for transformation 01

1 for all other transformations)

Variables M2 and R may be left blank if not required

C. Data Cards (optional-normally included): These cards supply the data for the problem to the computer. They must always be punched in FORMAT 12 F6.0 (i.e., twelve six-column fields per card; the last eight columns of each card may be used for identification). The data may be read in observation by observation or variable by variable, or the data for a particular problem may be taken from the immediately preceding problem, avoiding the need to read in any new data. Which of these options is adopted depends on the value of the parameter 'IN' punched on the control card (see Appendix 1).

D. Selection Cards (optional-normally included): No distinction is made, in reading in the data, between dependent and independent variables. Consequently it is possible to estimate successive equations with different dependent variables. For each equation which it is desired to estimate, one selection card must be punched in the following format:-

<u>Columns</u>	<u>Contents</u>
1	Option code for re-estimation of equation in homogeneous form (i.e., with intercept constrained to zero): 0 - This option not required 1 - This option is required
2	Option-code for print-out of table of residuals for this equation: 0 - Table of residuals not printed out 1 - Table of residuals is printed out 6 - This applies only to equations estimated in log form. As with 1 the table of residuals is printed out, and in addition, the anti-logs of the actual and predicted Y-values are printed.
3-4	Subscript of dependent variable for this equation.
5-6	Number of independent variables to be included as regressors in this equation.
7-8	} Subscripts of the independent variables required.
9-10	
11-12	
etc.	

This completes the list of cards which may be included in a single problem. The cards must always be read in the order shown. Of course, not all the cards mentioned will be required for every problem; as noted already, only the control card is obligatory for every problem. Thus, for example, many problems will not require any transformation of the input data, while problems which make use of data read in for a previous problem will not require any data cards. Note finally, that the present capacity of the program permits for each problem a maximum of fifty variables (including those read in and those transformed), fifty transformations, and ninety-nine equations.

5. Examples of Preparing Input for the Regression Program

Example 1: The first example is taken from [5], p.139. Four variables are read in, of which the fourth is the dependent; no transformations are required; and three equations are to be estimated using all ten observations. The cards required for this problem are listed in Appendix 4. Notice that two alternative ways of reading in the data are given: either observation by observation, in which case the control card and data cards are given by A and C; or variable by variable, corresponding to A' and C'. Notice also that all variables punched are right-justified, and that zeros may be left blank.

Example 2: This follows immediately after Example 1 and uses the same data, so "2" is punched in column 25 of the control card, and no new transformation or data cards are read in. Also the equations are re-estimated omitting the first and last observations, i.e. starting with observation 2 and ending with observation 9.

Example 3: The final example illustrates the flexibility of the data transformation option: only one variable is read in on cards, the Irish all items consumer price index, net of indirect taxation, from February 1958 to February 1972 (taken from [1] p.27); but a total of ten new variables are generated by the transformation cards shown. The reader should satisfy himself that these lead to the following variables:

- 1 - CPI, net of tax (as read in)
- 2 - logarithm of 1 to base ten
- 3 - 1 lagged
- 4 - 1 seasonally corrected
- 5 - percentage first difference of 4
- 6 - percentage first difference of 1
- 7 - time trend, with 1st quarter 1958 = 1.
- 8 - quadratic time trend
- 9 - seasonal dummy variable, 1 in first quarter, zero in all others
- 10 - seasonal dummy variable, 1 in second quarter, zero in all others
- 11 - seasonal dummy variable, 1 in third quarter, zero in all others

Note finally that for those transformations which take lags or first differences, the first observation on the new transformed variable is set equal to zero (for longer lags, the second or more observations may also be set equal to zero). In such circumstances, it is obviously meaningless to estimate equations over a set of observations including some which have been set equal to zero. Consequently, in the present example, the regression begins with the second observation.

6. Conclusion

It is hoped to add further options to the program in the near future. In the meantime the list already included should go some way towards facilitating applied econometric work. Finally, the author welcomes comments from users, concerning both improvements in the existing program and write-up, and possible additions which may be made to it.

Peter Neary

21 August 1972

References

1. Baker, T.J., and P. Neary: 'A Study of Consumer Prices, Part 1', in Quarterly Economic Commentary, E.S.R.I., March 1971.
2. Geary, R.C.,: 'Relative Efficiency of count of sign changes for assessing residual autoregression in least squares regression', Biometrika, vol. 57, no.1, 1970, pp. 123-127.
3. Haitovsky, Y.: 'Multicollinearity in Regression Analysis: Comment', Review of Economics and Statistics, Nov. 1969, pp. 486-489.
4. IBM Corporation: 'System/360 Scientific Subroutine Package, Version 3, Programmer's Manual', (5th edition: publicationGH 20-0205-4), New York 1970.
5. Johnston, J.: Econometric Methods, McGraw-Hill, 1963.
6. Leser, C.E.V.: 'Estimation of quasi-linear trend and seasonal variation', Journal of the American Statistical Association, vol. 58, 1963, pp.1033-1043.
7. Leser, C.E.V.: 'Seasonality in Irish Economic Statistics', E.S.R.I. Paper No. 26, 1965.
8. Longley, J.W.: 'An Appraisal of Least Squares Programs for the Electronic Computer from the Point of View of the User', Journal of the American Statistical Association, vol. 62, September 1967, pp. 819-841.
9. Mullet, G.M., and T.W. Murray: 'A New Method for Examining Rounding Error in Least Squares Regression Computer Programs', Journal of the American Statistical Association, vol. 66, September 1971, pp. 496-8.
10. Stewart, J.: 'A Method of Computing Homogeneous Least Squares Regression Equations', School of Social Sciences, New University of Ulster, Discussion Paper in Economics No. 4, July 1969.

APPENDIX 2

CODES FOR TRANSFORMATIONS

- 01 Addition of a series and a constant - $X(\text{MNEW}) = X(\text{MOLD}) + R$
02 Addition of two series, the second of which is multiplied by a constant - $X(\text{MNEW}) = X(\text{MOLD}) + X(\text{M2}) * R$
03 Multiplication of a series by a constant - $X(\text{MNEW}) = X(\text{MOLD}) * R$
04 Multiplication of one series by another series and a constant - $X(\text{MNEW}) = X(\text{MOLD}) * X(\text{M2}) * R$
05 Multiplication of one series by a constant, followed by division by another series - $X(\text{MNEW}) = X(\text{MOLD}) * R / (\text{M2})$
06 Log of a series to base e
07 Log of a series to base 10
08 Generation of a time trend (observation 'MOLD' set equal to one)
09 Generation of quarterly seasonal dummies (observation 'MOLD' taken as first quarter)
 $X(\text{MNEW}) = 1$ for all $(\text{MOLD} + (4 * N))$ observations
 $X(\text{MNEW} + 1) = 1$ for all $(\text{MOLD} + 1 + (4 * N))$ observations
 $X(\text{MNEW} + 2) = 1$ for all $(\text{MOLD} + 2 + (4 * N))$ observations
10 Change of subscript - $X(\text{MNEW}) = X(\text{MOLD})$
11 Series lagged one period - $X(\text{MNEW}) = \text{XLAG}(\text{MOLD}, 1)$
12 Series lagged two periods
13 Series lagged three periods
14 Series lagged four periods
15 First difference of series - $X(\text{MNEW}) = X(\text{MNEW}) - \text{XLAG}(\text{MOLD}, 1)$

16 Do., lagged one period
17 Do., lagged two periods
18 Do., lagged three periods
19 % first difference of series
 $X(\text{MNEW}) = X(\text{MOLD}) - \text{XLAG}(\text{MOLD}, 1) * 100.0 + \text{XLAG}(\text{MOLD}, 1)$
20 Do., lagged one period
21 Do., lagged two periods
22 Do., lagged three periods
23 Three quarter moving average of series
24 Do., lagged one period
25 Do., lagged two periods
26 % first differences, three quarter moving average of series
27 Do., lagged one period
28 Five quarter moving average of series
29 Raise a (positive) series to a real power: $X(\text{MNEW}) = X(\text{MOLD}) ** R$
30 Inverse of a series, multiplied by a constant - $X(\text{MNEW}) = (1 + X(\text{MOLD})) * R$
51 Seasonal correction of the quarterly variable 'MOLD', by means of Leser's quasi-linear trend method (for details see [6] and [7]. Observations over at least five full years on the variable must be available; and no zero or negative entries must be present. For this transformation, variables 'KCODE', 'MNEW' and 'MOLD' have their usual meanings, 'M2' is the number of the earliest observation which refers to the first quarter of a full year (if 'M2' is left blank the first observation is assumed to be the first quarter of a year), and 'R' is the year in which the first observation falls.

Appendix 3.

ERROR MESSAGES

No	Error	Message Printed Out	Action Taken
1	NTOT = 0	None	Job terminated
2	Either NTOT or MTRANS is greater than the capacity of the program (50)	Type 2 error	Job terminated
3	NLAST greater than NTOT	Type 3 error	Job continues, and assumes NLAST = NTOT
4	MTRANS = 0, but M \neq MTOT	Type 4 error	Job continues, and assumes M = MTOT
5	A parameter on the control card has a negative value	Type 5 error	Job continues, and substitutes a value of zero for the parameter in question.
6	Data matrix to be read in is too large for core (only possible when IN = 3, or seasonal corrections are requested)	Type 6 error	Job terminated
7	NS non-positive	Number of selections not specified.	Job continues, and assumes no selections are required for this problem.
8	A transformation has been requested whose code is not among those in Appendix 2.	Type 8 error	This transformation is ignored.
9	A transformation has been requested to generate a new variable with subscript MNEW, but MNEW exceeds both M and MTOT.	Type 9 error	This transformation is ignored
10	Transformation 51 (seasonal correction) has been requested for a variable which was not read in on cards	Type 10 error	Variable X(MOLD) is substituted for variable X(MNEW)
11	Transformation 51 (seasonal correction) has been requested for a variable on which less than five full years observations are available	Type 11 error	Same as for type 10 error
12	Restrictions on input variables to be transformed are violated, e.g., negative or zero observations are present in variables for which transformations 6,7, 29 or 51 were requested.	Observations - on input variable - violates restrictions for transformation -	For transformation 51 the transformation is ignored; for all other transformations a value of zero is inserted in the relevant observations on the new variables
13	In a particular selection the matrix to be inverted is singular	The matrix is singular	This selection is skipped
14	In a particular selection, the number of independent variables plus one is not less than the number of observations available.	Type 14 error	This selection is skipped



APPENDIX 4 : EXAMPLES 2 AND 3

PUNCHING INSTRUCTIONS

Name		Page		of		Application																																																																																																															
Date																																																																																																																					
1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	26	27	28	29	30	31	32	33	34	35	36	37	38	39	40	41	42	43	44	45	46	47	48	49	50	51	52	53	54	55	56	57	58	59	60	61	62	63	64	65	66	67	68	69	70	71	72	73	74	75	76	77	78	79	80																																						
A		EXMPL 2		1	0	4	3	2	9	2		1																																																																																																									
D		1	4	1	1																																																																																																																
		1	4	2	1	2																																																																																																															
		1	4	3	1	2	3																																																																																																														
A		EXMPL 3		5	7	1	1	1	2	1		8	3	1	1	1																																																																																																					
B		7	2	1																																																																																																																	
		1	1	3	1																																																																																																																
		1	5	5	4																																																																																																																
		5	1	4	1	1	1	9	5	8																																																																																																											
		1	5	6	1																																																																																																																
		8	7	1																																																																																																																	
		2	9	8	7	2	0																																																																																																														
		9	9	1																																																																																																																	
C		1	1	5	4	1	1	6	6	1	1	6	9	1	1	6	9	1	1	7	7	1	1	7	6	1	1	5	6	1	1	4	9	1	1	5	4	1	1	7	0	1	1	6	9	1	1	7	9	CP/NT	60	0																																																																	
		1	1	8	7	1	1	9	8	1	2	0	0	1	2	0	6	1	2	2	8	1	2	5	2	1	2	4	4	1	2	4	1	1	2	6	3	1	2	5	9	1	2	5	8	1	2	7	0	CP/NT	63	3																																																																	
		1	2	7	6	1	3	1	9	1	3	3	6	6	1	3	5	2	1	3	6	6	1	3	8	7	1	3	8	4	1	3	8	5	1	3	8	6	1	4	0	0	1	4	1	8	1	4	2	0	CP/NT	66	6																																																																
		1	4	2	2	1	4	4	1	1	4	4	2	1	4	5	2	1	4	8	4	1	5	0	1	1	5	0	5	1	5	1	8	1	5	6	3	1	5	8	3	1	6	0	3	1	6	2	6	CP/NT	69	9																																																																	
		1	6	4	6	1	6	8	2	1	7	1	4	1	7	5	1	1	7	8	2	1	8	2	9	1	8	6	8	1	9	1	0	1	9	5	8																																																																																
D		1	1	5	7	8	9	1	0	1	1																																																																																																										