REGEN-COMPUTER PROGRAM TO GENERATE MULTIVARIATE OBSERVATIONS FOR LINEAR REGRESSION EQUATIONS*

BY YOEL HAITOVSKY AND SIDNEY JACOBS

In the past few years the National Bureau has been giving increasing emphasis to research in econometrics and methodology. As part of this effort, the REGEN (REgression GENorator) computer program was developed. This program offers the statistician a method of investigating the sampling properties of estimators when analytical methods fail or when they become more costly than the moderate computer time needed for this program. The program uses a Monte Carlo technique which simulates the taking of a random sample of multivariate observations which satisfy a linear equation. The user of the program may specify the equation, and the mean, standard errors, distributions, and autocorrelations of the independent variables and of the error term, as well as the correlation between successive variables. The commonly used distributions are included in the program; moreover provision is made for the user to add his own distributions to the program. Features included in the program are: lagged variables, multiple time series, multicollinearity, errors of measurement (superimposed on independent variables), sampling from previously simulated populations or from a multiple time series, and aggregation of observations.

All results of the program can be saved either temporarily in the internal computer memory or permanently on cards or magnetic tape. This makes it possible to modify previously generated observations by any of the above-mentioned techniques. Another powerful facility that uses this retrievability of the data is the repetition option, which allows previous data to be reused, but in new equations or with new error terms. The authors show how this makes it possible to simulate simultaneous equation problems.

1. INTRODUCTION

Properties of estimators can be derived not only by analytical methods, but also through experimentation with models with known (or prespecified) structure and properties. The most commonly used of this class of analyses is the so-called Monte Carlo method. Its main use in statistical methodology is to investigate sampling properties of estimators when analytical methods fail or are too cumbersome.[1] When the statistician has this objective in mind he may simulate a "universe" by specifying its structure and parameters, and the distributions of the random variables appearing in the "universe." Then, the "universe" is sampled and the statistical technique under investigation is applied to the sample. By repeating the last step enough times, he can generate the distribution of the estimators under investigation, from which their sampling properties can be derived.

It is the belief of the authors that the potentials of the Monte Carlo analysis are not fully recognized by most statisticians. Many other statisticians are reluctant

* The authors would like to thank the reading committee: V. K. Chetty, Gregory Chow, Thomas Sargent, and especially its chairman Franklin Fisher for comments and suggestions that improved both the program and the paper. We are indebted to Mr. Barry Geller for his patience and care in checking the accuracy of test runs of the program and for preparing Section IV of this paper; and to Mrs. Virginia Meltzer for editorial assistance. Finally, we are grateful to John R. Meyer for his steady encouragement and support for the project.

[1] An economic justification for the use of Monte Carlo analysis is given by R. Sommers (1965): "A capital intensive approach to the small sample properties of various simultaneous equation estimators." *Econometrica* 33, 1-41. His argument is that high power analytical ability is becoming a scarce resource as compared to the availability of computer time. Thus, rational economic behavior would involve shifting towards more capital intensive methods.

to use it because it lacks the elegance and generality of analytical methods. The authors indeed recognize these limitations and the danger of making erroneous inferences by investigating only a narrow range of possible structures and parameters. Thus there is an additional burden on the user of the Monte Carlo methods: the need to investigate a wide range of parameters and many possible combinations when several parameters are involved. However, the authors believe that if appropriate computer programs are made readily available to the statistician, the relative ease of applying this "experimental" approach will more than compensate for the extra work involved in checking a wide range of possibilities.

The specific purpose of REGEN, the computer program described in this paper, is to generate multivariate observations which satisfy a linear equation. These observations may serve as either a sample or a universe for analyzing estimators of linear regressions. A dependent variable will be constructed as a linear combination of several independent variables plus an error term with specified distribution and parameters. (It will be shown in Section II that this procedure may be used also to construct systems of structural equations.) The generated data will be printed and punched on cards or saved on magnetic tapes, and will be ready for the application of the estimation technique under investigation. The program was made flexible enough so that a great variety of structures and a wide choice of parameters and distributions may be postulated.

The plan of the paper is as follows: a detailed description of the main program and the available options is presented in Section II, the methods used in generating the variables and the random number generators used for this purpose are described in Section III. Section IV contains a listing of the input used in a specific example and the output generated by it, and finally Section V contains the listing of the computer algorithm, which also includes the input instructions.

II. Description of the Effect of the Algorithm

Basic Regression Generation

The program generates random variables $X^1, \ldots, X^p, \varepsilon$ by sampling, in effect, from infinite populations. $X^1, \ldots, X^p$ represent the independent variables in a regression, while ε is an error variable whose mean is forced to zero. The user specifies the number of independent variables p, the number of observations n, and, for each random variable, the distribution, its mean and standard error,[2] as well as the correlation r_i between the populations corresponding to variables X^i and X^{i-1}, for $i = 2, 3, \ldots, p$. The uniform, Gaussian-normal, exponential, and Cauchy distributions are available in the program. Other distributions can easily be added to the program by the user. (For details, see the comments in the program listing.)

The user can also specify autocorrelation coefficients for each independent variable, that is, correlations between X_t^i and X_{t-1}^i for $i = 1, \ldots, p$. By doing this, the user generates values which simulate the observations of a time series. The

[2] The user is not, however, limited to distributions whose first and second moments are defined. Thus for the Cauchy distribution, the program interprets the two input moments as the center of symmetry and the interquartile range (the distance between the 25th and 75th percentile), respectively, of the specified population. (Samuel S. Wilks, *Mathematical Statistics*, 1962, pp. 255–256.)

number of observations of this time series is an integer l, which is provided by the user. This number is used to subdivide the n observations on each independent variable X^i and on ε into m autocorrelated series with l observations per series, where $m \times l = n$. Altogether there are $m \times (p + 1)$ series composed of m sets of $(p + 1)$-tuples. Thus the program simulates m observational units (e.g. individuals or families) by providing m multiple time series, each consisting of l multiple observations of $(p + 1)$ tuples. Each observational unit is linked through time by the autocorrelations, which are common to all units, but may be different for each variable. Applications are discussed in Sections 3.2 and 4 below. Notice that when both the correlation and autocorrelation options are jointly used, i.e., when autocorrelated series are to be intercorrelated, the user must choose compatible specifications (for conditions that must be satisfied cf. Section 3, Step 3).

In addition the user may request that variable X^i be lagged by k_i observations, $i = 1, \ldots, p$. These lagged variables will appear (in addition to $X^1, \ldots, X^p$ and ε) as $X^{p+1}, \ldots, X^{p+q}$ where q is the number of k_i that are greater than zero. (In the following discussion q if present has been absorbed into p for brevity.)

Once these p variables have been generated, the program then calculates the dependent variable Y using the formula

$$Y = \beta_0 + \sum_{i=1}^{p} \beta_i X^i + \varepsilon \tag{1}$$

where regression coefficients $\beta_1, \ldots, \beta_p$ and an intercept β_0 are supplied by the user, and each variable X^i as well as ε is a column vector with n entries.

However, the user may request the program to include among the independent variables a dependent variable Y lagged by s (>0) observations. We write Y_t for observation t of Y and X_t^i for observation t of X^i, $t = 1, \ldots, n$. Then, the user must also supply initial values $Y_{-1}, Y_{-2}, \ldots, Y_{-s}$ for the dependent variable, and coefficients $C_1, C_2, \ldots, C_s$ for the autoregressive part. Then for $t \geq 0$, equation (1) becomes

$$Y_t = \beta_0 + \sum_{k=1}^{s} C_k Y_{t-k} + \sum_{i=1}^{p} \beta_i X_t^i + \varepsilon. \tag{2}$$

By the above procedures the program generates a set of observations satisfying the user's specifications. This much of the procedure will be referred to as a basic regression generation. After a basic regression generation, the user can (a) request modifications of it, and/or (b) request another basic regression generation, and so on. The ouput from a basic regression generation, and from each modification requested, consists of the dependent variable Y, the independent variables X^i and the error variable ε, for each observation. The output can be printed, punched, and/or saved on binary tape at the user's option.

Modifications:

1. *Multicollinearity*

Adjoin one additional variable which is a specified linear combination of the independent variables $X^1, \ldots, X^p$:

$$X^{p+1} = \sum_{i=1}^{p} \alpha_i X^i.$$

Then recalculate the dependent variable Y, using specified regression coefficients which may be different from those that were used in the basic regression generation. This step may be repeated several times in succession to generate several additional variables that are linear combinations of the preceding variables. This modification, in conjunction with Modification 2 below, generates regressions that are useful in studying multicollinearity.

2. *Errors of Measurement*

For each variable X^i, $i = 1, \ldots, p$ obtained in the basic regression generation (or in the multicollinearity option), generate an error variable E^i whose distribution and standard error is specified by the user. Then superimpose this error on the variable: $\tilde{X}^i = X^i + E^i$. The user has the choice, in this modification, of recalculating or not recalculating the dependent variable Y. If Y is *not* recalculated, E^i can be regarded as the error of measurement in the variable X^i, whereas the ε variable mentioned earlier can be regarded as the error of measurement in the variable Y. (The user can also specify the mean and autocorrelation of each E^i, and the correlation between E^i and E^{i-1}, $i = 2, \ldots, p$, but all these are normally zero.)

3. *Sampling*

3.1. *Sampling form a single population.* The output discussed so far may be regarded as a random sample from an infinite population, or it may be itself regarded as the total population. In the latter case, the user can use the sampling option to request a random sample, of specified sample size, from this total population. (This is a special case of 3.2 below.) This can be done repeatedly to simulate repeated sampling from a fixed population.

3.2. *Sampling from a repeated population over time.* This modification considers the observations obtained by a basic regression generation, or by a previously executed modification, to be a succession of m observational units, each unit being traced over l time intervals (i.e., a time series of l observations). For each of the l time intervals, the program selects a random sample of the m observational units and outputs the observations for these units only. The user can request either a new sample for each time period or the same sample for all the time periods, i.e., a panel. The user specifies l, m, and the sample size.

4. *Aggregation*

This modification aggregates observations by summing for each time interval over the observations obtained from a basic regression generation or from a previous modification. This is most likely to be of interest when the observations were obtained from a basic regression generation using the time series option.

Modifications 3 and 4 can be usefully combined, and applied to a Monte Carlo study of regression parameters estimated by pooling time series and cross-section data.[3] This is done by generating time series for each individual in a population.

[3] Cf. J. Durbin, "A Note on Regression when there is Extraneous Information about one of the Coefficients," *Journal of the American Statistical Association*, 48 (1958), 799–808; V. K. Chetty, "Pooling of Time Series and Cross-Section Data," *Econometrica*, Vol. 36, No. 2 (1968), 279–290.

Aggregation across the population in a given "time interval" results in aggregative time series data. Sampling within a "time interval" results in cross sectional samples.

5. *Repetition of Data*

5.1. In this modification the program generates an error vector only, and then recalculates the dependent variable in accordance with equation (1), using previously generated values of the independent variables. (These repeated values of the independent variables may be already in core or may be read into core from a binary tape that was either created by this program or obtained from another source.) If the regression coefficients and the specifications of the error variable are left the same for each repetition, the process can be interpreted as simulating repeated samples drawn from the same distribution. However, one can also change the error specification and regression coefficients.

5.2. *Simultaneous equations.* Another application of this option, in which the regression coefficients are different for each repetition, is to the so-called simultaneous equation problem in which one has a system of equations

$$Y\Gamma = XB + U \tag{3}$$

in which Γ is a nonsingular $p \times p$ matrix, Y and U are each $t \times p$ matrices, while X is $t \times q$ and B is $q \times p$. Here each column of the matrix U represents an error term. If we multiply both sides of (3) by Γ^{-1}, we obtain an equivalent system of equations called the reduced form:

$$Y = X\pi + U\Lambda \tag{4}$$

where $\pi = B\Gamma^{-1}$ and $\Lambda = \Gamma^{-1}$ are respectively $q \times p$ and $p \times p$ matrices.

To generate this problem with our program, we regard (4) as p distinct problems, where each problem has the form:

Column i of lefthand side of (4) = column i of righthand side of (4) for

$$i = 1, \ldots, p,$$

or formally, problem number i is

$$Y^i = X^1\pi_1^i + \ldots + X^q\pi_q^i + U^1\Lambda_1^i + \ldots + U^p\Lambda_p^i$$

where Y^i, X^j, U^k are the ith, jth, and kth $t \times 1$ column vectors of matrices Y, X, and U respectively ($i = 1, \ldots, p; j = 1, \ldots, q; k = 1, \ldots, p$) and π_j^i and Λ_j^i are the entries in column i, row j of matrices π, Λ respectively. Thus problem number i is just an ordinary regression (see equation (1)) with dependent variable $Y = Y^i$, constant $\beta_0 = 0$, $q + p$ independent variables $X^1, X^2, \ldots, X^q, U^1, U^2, \ldots, U^p$, and error term $\varepsilon = 0$, and regression coefficients

$$\beta_i = \pi_1^i, \beta_2 = \pi_2^i, \ldots, \beta_q = \pi_q^i, \beta_{q+1} = \Lambda_1^i, \ldots, \beta_{q+p} = \Lambda_p^i. \tag{5}$$

If one is interested in including an intercept in (3), one should interpret X^1 as a $t \times 1$ vector of ones. In this case π_1^i should be input as the (possibly nonzero) intercept β_0 for the i-th problem. To generate all the data for the reduced form of the simultaneous equation problem, one runs problem number 1 with its beta

coefficients (equation (5) with $i = 1$), obtaining $Y^1, X^1, \ldots, X^q, U^1, \ldots, U^p$. Then one runs successively problems 2, 3, ..., p using the repetition option in each case, obtaining $Y^2, \ldots, Y^p$. For each of the p problems one sets the error vector ε equal to zero, by specifying its standard error to be zero.

Any of these p problems can also include a lagged dependent variable (see equation (2)). In particular, if one wants the same lagged variable to be used in several of the p problems, one can arrange for the lagged variable to be Y^1, generate Y^1_{-1}, Y^1_{-2}, etc. in problem 1, and repeat this lagged variable in any of problems 2, 3, ..., p.

Finally, we illustrate how the repetition of data option can be used to generate a system of equations in which Y^2 lagged occurs in the 1st equation and Y^1 lagged occurs in the 2nd equation. For simplicity, assume that there are only two equations in all, and that they are already in reduced form:

$$Y^1 = \beta_0 + \sum_{i=1}^{p} \beta_i X^i + \beta_{p+1} Y^2_{-1} + \lambda_{11} U^1 + \lambda_{12} U^2 \tag{6}$$

$$Y^2 = \gamma_0 + \sum_{i=1}^{p} \gamma_i X^i + \gamma_{p+1} Y^1_{-1} + \lambda_{21} U^1 + \lambda_{22} U^2. \tag{7}$$

If we lag equation (7) by one observation we obtain an expression for Y^2_{-1} which can be substituted into equation (6), giving

$$\begin{aligned} Y^1 = \beta_0 + (\beta_{p+1}\gamma_0) + \sum_{i=1}^{p} \beta_i X^i + \lambda_{11} U^1 + \lambda_{12} U^2 + \sum_{i=1}^{p} \beta_{p+1}\gamma_i X^i_{-1} \\ + \beta_{p+1}\lambda_{21} U^1_{-1} + \beta_{p+1}\lambda_{22} U^2_{-1} + \beta_{p+1}\gamma_{p+1} Y^1_{-2}. \end{aligned} \tag{6'}$$

The pair of simultaneous equations (6'), (7) is equivalent to the given pair (6), (7), and moreover can be generated by REGEN, because Y^2 has been eliminated in (6'). One generates equation (6') as problem 1, in which one also generates Y^1_{-1} (even though it does not occur explicitly in the equation) along with the other lagged variables. Then, using the repetition of data option, one passes Y^1_{-1} on to equation (7), which is generated as problem 2.

III. Method of Generating Variables

Once the p independent variables have been generated, the remaining variables (namely the dependent variable, lagged dependent and independent variables), samples, and aggregates are generated in accordance with the formulae and methods described in Section II. It remains to describe, in the present section, the method of generating the p independent variables. The following discussion also applies to the generation of the error variables E^j used in Modification 2 (see Section II).

For each distribution included in the program, there is a generator, i.e., a subroutine that generates random numbers from that distribution. The mean and standard error of the population from which the generator draws its random numbers is dependent on the subroutine: when referring to the generator specified for the j-th variable we will call them μ_j and σ_j respectively. For each variable

$X^j, j = 1, \ldots, p$ (and for ε, which the program considers as X^{p+1}) the program invokes the generator for its distribution n times, obtaining random numbers $X_{ij}, i = 1, \ldots, n$. We denote the observed mean of these n numbers by M_j. (M_j will be statistically "near" to μ_j.)

As is well known all the statistics computed from the sample will differ from the corresponding population specifications. The deviation will depend on the sample size and on the population specifications. The deviation will be particularly pronounced when small correlation coefficients are specified (see R. A. Fisher, *Statistical Methods for Research Workers*, 1925, pp. 81–84) and when large variances (relative to the same size) are requested. The program is designed to reflect this situation, so that the observed statistics will be close, but not necessarily equal to the user's specifications.

These quantities X_{ij} undergo several operations whose effect is to transform the column vector X_{ij} so that it meets the user's specifications for the variable X^j. We use the following notation for the user's specifications.

XM_j	mean of j-th variable
S_j	standard error of j-th variable
A_j	autocorrelation of j-th variable
R_j	correlation between X^{j-1} and X^j

The steps to obtain the variable X^j are as follows:

Step. 1. Calculate the observed mean $\bar{X}_j$ of X^j:

$$\bar{X}_j = XM_j + (M_j - \mu_j)\frac{S_j}{\sigma_j}.$$

Step 2. Replace X_{ij} by a normalized variable whose observed mean is 0, and whose expected standard error is 1.0:

$$X_{ij} = (X_{ij} - M_j)/\sigma_j.$$

Step 3. If the time series option is used, an autocorrelation equal to A_j is specified for X^j. The interdependence between autocorrelations and intercorrelations is expressed by the restriction

$$-[1.0 - R_j^2(1.0 + A_{j-1})] < A_j < 1.0 - R_j^2(1.0 - A_{j-1})$$

for $j = 2, \ldots, p$. If the specified A_j does not meet this restriction, it is equated to the nearer of the two bounds.

Moreover, because of this interdependence, one finds that one must impose an autocorrelation equal to A_j' instead of A_j, where

$$A_1' = A_1.$$

$$A_j' = (A_j - R_j^2 A_{j-1})/(1 - R_j^2). \qquad j = 2, \ldots, p.$$

The variable X^j with expected autocorrelation A_j', but with expected mean and standard error unchanged from Step 2, is produced by the replacement:

$$X_{1j} \text{ unchanged}$$

$$X_{ij} = X_{i-1,j}A_j' + X_{ij}\sqrt{1.0 - (A_j')^2}. \qquad i = 2, \ldots, n.$$

Step 4. Impose an expected correlation R_j between X^j and X^{j-1} by the replacement

$$X_{ij} = X_{i,j-1}R_j + X_{ij}\sqrt{1 - R_j^2}$$

successively for $j = 2, \ldots, p$. This also has the effect of changing the expected autocorrelation of X^j from the value A_j' of Step 3, back to A_j as specified, but without disturbing the mean and standard error of Step 3.

Step 5. Change the expected mean and standard error of X^j to the specified values by the replacement

$$X_{ij} = X_{ij}S_j + \bar{X}_j.$$

The simplest illustration of the above operations is for the case where all the A_j's and R_j's are zero. Then the net effect of Steps 1–5 is the transformation

$$X_{ij} = XM_j + (X_{ij} - \mu_j)\frac{S_j}{\sigma_j}.$$

Remarks on the Random Number Generators

In the following paragraphs, the particular random number generators implemented in the program are identified. We emphasize, however, that the user can incorporate his own generators into the program either in place of or in addition to those presently in the program. For each generator, one supplies to the main program the mean and standard error of the distribution, one call statement to the subroutine, and, if appropriate, statements to supply a starting random number or retrieve the final random number.

The uniform random number generator used by the program is RANNO (Harvard Computing Center), which employs the power residue method to generate random numbers between 0 and 1. It allows the program to supply the starting random number, which is to be specified by the user. In addition to its use in producing random variables, RANNO is also used to select sample observations in Modification 3, Section II.

The normal random number generator is GAUS, which applies the central limit theorem to 12 uniform random numbers obtained from RANNO to generate one normal number. Subroutine GAUS thus indirectly uses the same starting random number as RANNO.

The random numbers from the exponential and Cauchy distributions are obtained from entry points EXPRN and CAUCHY respectively of a subroutine ORMC, which calls a subroutine FLOAT. The subroutines for these two distributions were originally coded for the IBM 704 by R. E. Coveyou and J. G. Sullivan (SHARE distribution No. 743) and have been slightly modified for use on the IBM 7094. A version for System 360, with different random number routines will be available shortly.

IV. Sample Input and Output

A sample of the input options and variable parameters (not in input format) as well as the resulting output generated from REGEN are shown in Tables A and B.

TABLE A

INPUT OPTIONS FOR THE REGRESSION GENERATION PROGRAM

No. of Observations = 20
No. of Variables = 4

NEW	SAVE	PUN	COV	LC	EM	SAMP	FIXS
0	0	0	1	0	0	0	0

CS	REX	TS	DB	TAPE	NOP	NOREP	LTS
0	0	1	1	0	5	0	5

Starting Random No. = 13579

Distribution for Variable	1	2	3	4	5	6	7	8	9	10
	1	2	3	4	2	0	0	0	0	0

Dependent Variable Lagged by 0 Observations

Independent Variable No.	1	2	3	4	5	6	7	8	9	Lagged by
	0	0	0	0	0	0	0	0	0	Observations

Maximum Lag of Indep. Variables is 0
No. of Lagged Indep. Variables is 0

Specified Means of Independent Variables
100.000 50.000 1.000 −20.000
Specified Standard Errors of Independent Variables (Error Term Last)
10.000 1.000 5.000 10.000 100.000
Specified Auto-correlations of Independent Variables
0.500 0.595 0.741 0.900 0.000
Specified Correlations Between Successive Variables
0.000 0.900 0.800 0.600
Regression Coefficients (Intercept Last)
2.000 1.000 20.000 5.000 45.000

respectively. Appendix I consists of a listing of the first part of the program, while Appendix II shows the input for the sample problem. The user specifies the following (see Table A):

(a) 20 observations on four independent variables are to be generated (NEW = 0);
(b) The observations shall be neither saved on binary tape (SAVE = 0) nor punched on cards (PUN = 0);
(c) Both the covariance and the correlation matrices of the independent variables are to be printed (COV = 1);
(d) No variable which is a linear combination of independent variables is created (LC = 0) and no errors of measurement are superimposed on any of the independent variables (EM = 0);
(e) No previously generated time series data in core are to be used (SAMP = FIXS = 0) and thus, no aggregation of previously generated time series (CS = 0);
(f) No generation of the error for repeated use of the data (REX = 0);
(g) Autocorrelated variables are generated (TS = 1);
(h) The unadjusted random numbers as generated by the random number generator are printed (DB = 1), for the first five observations (of the total 20) that are to be printed (NOP = 5);
(i) No printing is suppressed (NOREP = 0) except for the last 15 observations, and no variables (independent, error, or dependent) are obtained from binary tape (TAPE = 0); and
(j) Five time intervals of time series data are requested (LTS = 5).

The program also prints the number from which the random number generator begins generating random numbers (13579) after which the distribution for each variable is shown. As can be seen from Table A, variable 1 has a normal distribution (code 1), variable 2 has a uniform distribution (2), variable 3 has an exponential distribution (3), variable 4 has a Cauchy distribution (4) and the error variable (variable 5) has a uniform distribution (2). There are no lagged variables in the sample.

The program then prints the parameters of the independent variables specified by the user. For example, variable 1 has a mean of 100 and a standard error of 10 while the error term (printed last) has a mean of zero and a standard error of 100. The auto-correlations of the independent variables as well as the specified correlations between successive variables are also printed in Table A. Moreover, the user-specified regression coefficients, which are used to derive the dependent variable, are printed (with the intercept term last).

The generated output is shown in Table B. The unadjusted random numbers as generated by the respective random number generators for each of the four independent variables and for the error term are shown for the first five observations with respective means, variances and standard errors. For example, the first five observations on variable 1 have mean = −0.32915, variance = 0.72072 and standard error = 0.84895. These random numbers are then modified by the input specifications outlined above and the resulting numbers, which conform to the input specifications, are shown in the X-matrix of Table B. The Y-vector of dependent variables (to the left of this X-matrix) is derived by excluding the last column vector of error terms from the X-matrix, multiplying the remaining X-matrix by the vector of pre-specified regression coefficients (excluding the intercept), adding the column vector of error terms to this product and finally adding the pre-specified intercept value to each element of the resulting vector. For example, the first element in the Y vector is derived by first adding the products of the elements of the first row vector of the X-matrix (excluding the last term) and the respective elements of the column vector of pre-specified regression coefficients (excluding the intercept) and then adding this sum to the sum of the prespecified intercept value and the first element of the error vector as follows:

$$-165.23779 = 2(93.64462) + 1(48.83626) + 20(-4.39606) + 5(-38.27237) + 45 - 167.08019$$

Thus, the X-matrix consists of the generated independent variables and the error terms (printed last) and the Y-vector is the vector of dependent variables generated from this X-matrix and the vector of pre-specified regression parameters.

The program then prints the observed means, variances and standard errors of each of the output variables as well as the covariance and correlation matrices of each of the four independent variables and the auto-correlation coefficients. For example, variable 1 has pre-specified mean equal to 100 and pre-specified standard error equal to 10 (see Table A) while the variable 1 produced by the program has mean equal to 95.78670, variance equal to 29.96899 and standard error equal to 5.47439. It should be noted that the diagonal elements of the covariance matrix indicate the variance of the output variables, while the off-diagonal elements show

their covariances. For example, variable 1 has variance equal to 29.96924 while the covariance between output variable 1 and output variable 2 is shown to be equal to 2.69739 (see Table B). The auto-correlation coefficients of the output variables are shown at the bottom of Table B.

TABLE B

OUTPUT OF THE REGEN PROGRAM BASED ON THE INPUT SPECIFICATIONS OF TABLE A

Regression Generation

Unadjusted Random Numbers

−0.63554	0.05784	0.13741	−0.81710	0.01768
−0.88170	0.97590	0.15926	−0.41353	0.03036
0.72417	0.33483	4.85599	−0.41580	0.02304
−1.24956	0.22587	0.54992	−0.89915	0.86497
0.39685	0.34075	0.00791	5.20690	0.98249
Mean				
−0.32915	0.38724	1.14210	0.53226	0.38371
Variance				
0.72072	0.12143	4.35134	6.87883	0.24477
Sigma				
0.84895	0.34853	2.08599	2.62275	0.49474

Regression Generation

Observational Unit No. 1

Y	X Matrix				
−165.23779	93.64462	48.83626	−4.39606	−38.27237	−167.08019
−207.85787	90.39138	48.54377	−5.56651	−41.63328	−162.68775
−79.44145	102.67194	49.64899	−1.14369	−38.26707	−165.22511
71.60774	91.71927	48.66317	−5.08730	−46.03539	126.42891
268.26692	100.50128	49.45353	−1.92645	−31.15973	167.13852

Regression Generation

Observed Means, Variances, and Standard Errors

Y	X Variables				
Mean					
−22.53248	95.78670	49.02914	−3.62400	−39.07357	−40.28513
Variance					
37843.89648	29.96899	0.24283	3.88606	29.73509	29371.88159
Sigma					
194.53508	5.47439	0.49278	1.97131	5.45299	171.38227

Regression Generation

Covariance			
29.96924	2.69739	10.79171	20.81854
2.69739	0.24786	0.97127	1.87347
10.79171	0.97127	3.88606	7.49385
20.81954	1.87347	7.49385	29.73514
Correlations			
1.00000	0.99990	1.00000	0.69740
0.99990	1.00000	0.99985	0.69721
1.00000	0.99985	1.00000	0.69713
0.69740	0.69721	0.69713	1.00000
Auto-correlations			
−0.68523	−0.68441	−0.68482	−0.95826

Copies of the REGEN program (decks or print-outs) and related documentation are available at marginal cost from the National Bureau of Economic Research. Contact Charlotte Boschan, Chief of Data Processing.

The Hebrew University Jerusalem and
National Bureau of Economic Research
National Bureau of Economic Research

APPENDIX I

Listing of the First Part of REGEN Program

INPUT INSTRUCTIONS.
REFER TO STATEMENTS 125, 180, 181, 195 THRU 215, AND 345.
PROGRAM READS AN OPTION CARD AND ANY SPECIFICATION CARDS REQUIRED. IT THEN GENERATES DATA SPECIFIED BY THESE CARDS - I.E. A BASIC REGRESSION GENERATION (0 IN COLUMN 9) OR ONE OF THE MODIFICATIONS DEFINED IN COLUMNS 13-19 (1 IN COLUMN 9). THEN THE PROGRAM RECYCLES TO STATEMENT 125 TO READ ANOTHER OPTION CARD, AND SO ON. EXECUTION IS TERMINATED ON READING AN OPTION CARD WITH BLANKS IN COLUMNS 1-4.

(I) OPTION CARD

EF IS THE COLUMN NO. OF THE RIGHT-MOST COLUMN OF THE FIELD.

EF	NAME	DEFINITION
04	NOBS	NO. OF OBSERVATIONS. NCBS+MLX MUST NOT EXCEED 1700.
06	NVAR	NO. OF INDEPENDENT VARIABLES X. IT DOES NOT INCLUDE LAGGED INDEPENDENT VARIABLES, LAGGED OR UNLAGGED DEPENDENT VARIABLES Y, NOR THE ERROR VARIABLE EPSILON. NVAR+NLX MUST NOT EXCEED 9 BUT NVAR MUST BE AT LEAST ONE.
08	INEW	=0. REQUEST BASIC REGRESSION GENERATION (BRG). SPECIFICATION CARDS FOR XM, V, GAMMA IF ITS=1, W, AND BETA ARE REQUIRED. =1. REQUESTS MODIFICATION (MOD.) OF PREVIOUS BRG OR OF PREVIOUS MOD. SPECIFIC MOD. S ARE DEFINED IN COLUMNS 13-19.
10	ISAVE	=2. SAVE X, LAGGED X, LAGGED Y, EPSILON, AND UNLAGGED Y, IN THAT ORDER, ON BINARY TAPE 9. THIS TAPE IS REWOUND BEFORE WRITING. =1. SAVE THE OBSERVATIONS AS ABOVE, BUT OMIT EPSILON. =0. DO NOT SAVE THE OBSERVATIONS ON BINARY TAPE.
11	IPUN	=2. PUNCH SERIAL NO. OF OBSERVATION, X, LAGGED X, LAGGED Y, EPSILON, AND UNLAGGED Y, IN THAT ORDER, ON CARDS (USING SYSTEM PUNCH TAPE 7) IN FORMAT (14,8F9.3 (4X8F9.3)). =1. PUNCH THE OBSERVATIONS AS ABOVE, BUT OMIT EPSILON. =0. DO NOT PUNCH OBSERVATIONS.
12	ICOV	=1. PRINT COVARIANCE AND CORRELATION MATRICES FOR X VARIABLES (UNLAGGED AND LAGGED). =0. OMIT THIS PRINTOUT.
13	ILC	=1. MULTICOLLINEARITY. GIVEN P X-VARIABLES EITHER ALREADY IN CORE OR READ INTO CORE FROM BINARY TAPE 9 (SEE ISAVE=2 AND ITAPE=2 BELOW), GENERATE A (P+1)ST X-VARIABLE AS A LINEAR COMBINATION OF THE OTHERS. THIS MOD. REQUIRES SPECIFICATION CARDS FOR ALPHA AND BETA ONLY. NOTE - USE NVAR=P+1 FOR THIS MOD. =0. IGNORE.

14	IEM	=1. ERRORS OF MEASUREMENT. GIVEN NVAR PREVIOUSLY GENERATED VARIABLES, AND EPSILON AND Y, SUPERIMPOSE AN ERROR VARIABLE ON EACH OF THE NVAR VARIABLES, BUT LEAVE EPSILON AND Y UNCHANGED. THE PREVIOUSLY GENERATED DATA MUST BE ON BINARY TAPE 9, AND MAY HAVE BEEN CREATED BY A PREVIOUS BRG OR MOD. (SEE ISAVE=2 BELOW), OR MAY BE FROM SOME OTHER SOURCE. SPECIF. CARDS DEFINE THE SPECIFICATIONS OF THE ERROR VARIABLES. SPECIF. CARDS FOR XM, V, GAMMA IF ITS=1, AND W ARE REQUIRED. =0. IGNORE.
15	ISAMP	SAMPLING. IF GREATER THAN 0, ISAMP IS THE SAMPLE SIZE. FOR EACH TIME INTERVAL, RANDOMLY SELECT ISAMP OBSERVATIONAL UNITS, AND OUTPUT THE OBSERVATIONS FOR THESE UNITS ONLY. (THE NO. OF TIME INTERVALS IS LTS, Q.V.) USES PREVIOUSLY GENERATED TIME SERIES DATA ALREADY IN CORE OR READ INTO CORE FROM BINARY TAPE 9 (SEE ISAVE AND ITAPE BELOW). NO SPECIFICATION CARDS REQUIRED. =0. IGNORE.
17	IFXS	USED IN CONJUNCTION WITH ISAMP GREATER THAN 0. =0. FOR EACH TIME INTERVAL, SAMPLE THE OBSERVATIONAL UNITS INDEPENDENTLY. =1. FOR EACH TIME INTERVAL, USE THE SAME SAMPLE OF OBSERVATIONAL UNITS.
18	ICS	=1. AGGREGATION. AGGREGATE PREVIOUSLY GENERATED TIME SERIES DATA. (THE NO. OF TIME INTERVALS IN THE TIME SERIES IS LTS, Q.V.) NO SPECIFICATION CARDS REQUIRED. =0. IGNORE.
19	IREX	=1. REPETITION OF DATA. GENERATE AN ERROR VARIABLE EPSILON ONLY. USING PREVIOUSLY GENERATED INDEPENDENT VARIABLES. RECALCULATE THE DEPENDENT VARIABLE Y. SPECIFICATION CARDS FOR V, AND FOR GAMMA IF ITS=1 ARE NEEDED, IN WHICH ONLY EPSILON IS DEFINED, IN THE FIRST FIELD. THESE CARDS ARE FOLLOWED BY THE SPECIFICATION CARD FOR BETA, IN WHICH ALL NVAR + 1 VALUES OF BETA ARE GIVEN. =0. IGNORE.
20	ITS	=1. GENERATE AUTOCORRELATED VARIABLES FOR TIME SERIES DATA. THE NO. OF TIME INTERVALS IS LTS, Q.V. THIS OPTION IS AVAILABLE IN CONJUNCTION WITH BRG (INEW=0), ERROR OF MEASUREMENT (INEW=1 AND IEM=1), AND REPETITION OF DATA (INEW=1 AND IREX=1). THE PROGRAM GENERATES A COMPLETE SET OF LTS OBSERVATIONS ON THE AUTOCORRELATED VARIABLES FOR EACH OBSERVATIONAL UNIT SUCCESSIVELY. THE NO. OF OBSERVATIONAL UNITS IS GIVEN BY NOU=NOBS LTS. THE AUTOCORRELATIONS ARE SPECIFIED ON THE CARD FOR GAMMA, Q.V. =0. IGNORE.
21	IDB	=1. PRINT UNADJUSTED RANDOM NUMBERS AS GENERATED BY THE RANDOM NUMBER GENERATOR SUBROUTINES. FOR ERROR OF MEASUREMENT (IEM=1), ALSO PRINT THE MATRIX OF ERRORS. =0. IGNORE.
22	ITAPE	=2. OBTAIN THE INDEPENDENT VARIABLES, THE ERROR VARIABLE, AND THE DEPENDENT VARIABLE FROM BINARY TAPE 9 FOR FURTHER PROCESSING BY A MODIFICATION (INEW=1). THIS OPTION IS NOT AVAILABLE FOR A BRG (INEW = 0), AND

SHOULD NOT BE USED FOR THE ERROR OF MEASUREMENT MOD., WHICH AUTOMATICALLY OBTAINS DATA (INCLUDING THE ERROR AVAILABLE FROM BINARY TAPE 9)

=1. SAME AS ITAPE=2 ABOVE, BUT IT IS ASSUMED THAT THE ERROR VARIABLE IS NOT PRESENT ON THE TAPE.
=0. IGNORE.

31-40 IRN — AN ODD INTEGER OF 10 DIGITS OR LESS TO BE USED AS THE STARTING FIXED POINT QUANTITY FOR SUBROUTINES RANNO AND GAUS (UNIFORM AND NORMAL RANDOM NUMBER GENERATORS RESP.).
ON THE 2ND AND LATER OPTION CARDS, IRN=0 CAN BE USED TO SIGNAL THE PROGRAM TO CONTINUE WITH THE NEXT AVAILABLE STARTING RANDOM NO. AS SAVED FROM THE LAST PRECEDING BRG OR MOD. THE PRINTED OUTPUT SHOWS THE ACTUAL STARTING NO. USED, SO THAT THE USER CAN CONTINUE A SERIES OF EXPERIMENTS FROM WHERE HE LEFT OFF.

41-50 NUDIS(1) THRU NUDIS(10) — THIS FIELD IS IGNORED IN THE MULTICOLLINEARITY AND AGGREGATION MODS. WHICH DO NOT USE RANDOM NUMBERS
IN ONE-COLUMN FIELDS, SPECIFYING THE DISTRIBUTIONS OF X 1, X 2, ..., X-NVAR, AND EPSILON RESPECTIVELY.
THE CODES FOR THE DISTRIBUTIONS ARE
1 NORMAL
2 UNIFORM
3 EXPONENTIAL
4 CAUCHY
THE MEANS AND STANDARD ERRORS OF THESE DISTRIBUTIONS ARE GIVEN IN THE LISTING STARTING AT STATEMENT 120.
NOTE - IN THE ERROR OF MEAS. MOD., SUPPLY A CODE OF 2 FOR EPSILON EVEN THOUGH EPSILON IS NOT AFFECTED BY THIS MOD.
NOTE - IN THE REPETITION MOD., ONLY THE CODE FOR EPSILON, NAMELY NUDIS(NVAR+1), IS USED.

54 NOBSP — NUMBER OF OBSERVATIONS TO PRINT (OF EACH OBSERVATIONAL UNIT, IF ITS=1). IF 0, PROGRAM PRINTS ALL NOBS OBSERVATIONS (OR ALL LTS OBSERVATIONS OF EACH OBS. UNIT OF TIME SERIES DATA IF ITS=1).
IF NEGATIVE, NO OBSERVATIONS ARE PRINTED.

58 NOREP — =+1. SUPPRESS THE PRINTING OF MEANS, VARIANCES AND STANDARD ERRORS OF VARIABLES (AND OF AUTOCORRELATIONS IF ITS=1).
=-1. IN ADDITION SUPPRESS THE PRINTING OF THE REPORT OF INPUT. THIS SUPPRESSES ALL PRINTED OUPUT EXCEPT WHAT MIGHT BE REQUESTED BY THE SETTING OF IDB AND NOBSP.
=0. IGNORE.

62 LTS — NUMBER OF TIME INTERVALS OF TIME SERIES DATA.
NOTE- UNDER THE SAMPLING OR AGGREGATION MODS, AN INPUT VALUE OF LTS=0 IS RESET BY THE PROGRAM TO LTS=1, I.E. ONE SAMPLE OR ONE AGGREGATE OBSERVATION ONLY, RESP., IS TAKEN FROM THE ENTIRE SET OF OBSERVATIONS.

63 IQ — AN INTEGER BETWEEN 1 AND 8 INCLUSIVE, USED BY THE MULTICOLLINEARITY MOD. ONLY THE FIRST IQ OF THE P=NVAR-1 PREVIOUSLY GENERATED INDEP. VARIABLES WILL BE USED IN OBTAINING A LINEAR COMBINATION.

64 LY — NUMBER OF TIME INTERVALS BY WHICH THE DEPENDENT VARIABLE IS TO BE LAGGED. LY LAGGED DEPENDENT VARIABLES WILL BE OUTPUT WITH LAGS OF 1, 2, ..., LY RESPECTIVELY.
IF LY IS GREATER THAN 0, SPECIFICATION CARDS FOR BY AND YIN ARE REQUIRED.
=0. NO LAGGED DEPENDENT VARIABLES.

65–72	LX(1) THRU LX(8)	FOR EACH I = 1, 2, . . . , NVAR, FOR WHICH LX(I) IS NON-0, GENERATE I-TH INDEP. VARIABLE LAGGED BY LX(I) INTERVALS AS AN ADDITIONAL INDEP. VARIABLE.
74	MLX	MAXIMUM OF LX(I)'S ABOVE.
75	NLX	NO. OF LAGGED INDEP VARIABLES (= NO. OF NON-0 LX(I)'S ABOVE).

(II) SPECIFICATION CARDS

ALL THE CARDS HAVE FORMAT (10F8.3).
EACH ITEM IN THE NAME COLUMN IS A VECTOR PUNCHED ON A SEPARATE CARD.

NAME	MEANING
BY	COEFFICIENTS OF AUTOREGRESSION. BY(1), BY(2), . . . , BY(LY) REFER TO Y − 1, Y − 2, . . . , Y − LY RESPECTIVELY.
YIN	INITIAL VALUES OF DEPENDENT VARIABLE. Y − 1, Y − 2, . . . , Y − LY IN THAT ORDER. BY AND YIN ARE INPUT ONLY WHEN LY IS GREATER THAN 0.
XM	SPECIFIED MEANS OF INDEPENDENT VARIABLES. XM(I) IS THE SPECIFIED MEAN OF THE I-TH INDEP. VARIABLE.
V	SPECIFIED STANDARD ERRORS OF INDEPENDENT VARIABLES. V(NVAR + 1) IS THE SPECIFIED STD. ERROR OF EPSILON.
GAMMA	SPECIFIED AUTOCORRELATIONS OF INDEPENDENT VARIABLES AND OF EPSILON. INCLUDE THIS CARD ONLY WHEN ITS = 1.
W	SPECIFIED INTERCORRELATIONS BETWEEN SUCCESSIVE INDEPENDENT VARIABLES. W(1) IS IGNORED. FOR I = 2, . . . , NVAR, W(I) IS THE SPECIFIED CORRELATION BETWEEN THE (I − 1)ST AND I-TH INDEPENDENT VARIABLE.
BETA	SPECIFIED REGRESSION COEFFICIENTS FOR THE INDEPENDENT VARIABLES. IF NLX IS GREATER THAN 0, COEFFICIENTS BETA(NVAR + 1), . . . , BETA(NVAR + NLX) ARE INPUT FOR THE LAGGED INDEP. VARIABLES. IN ANY CASE THE INTERCEPT OR CONSTANT TERM IS INPUT AS THE LAST BETA, NAMELY BETA(NVAR + NLX + 1).
ALPHA	COEFFICIENTS FOR THE LINEAR COMBINATION GENERATED BY THE MULTICOLLINEARITY MOD. ONLY ALPHA(1) THRU ALPHA(IQ) ARE USED.

Appendix II

Input For Sample Problem

```
20   40001000000011000000000000001357912342000000000000000005000000000000
100.      50.       1.      −20.
10.       1.        5.       10.      100.
 .5        .6        .8        .9        .0
                     .9        .8        .6
 2.       1.        20.       5.       45.
20   4120100 3000010000000000000909999900000000000000000000005000000000000
15   4100100000100120000000000000000000000000300000000000000005000000000000
 5   40001000900011000000090000001357912342000000000000000005021302000003 3
 .5       −.5
50.       40.
100.      50.       1.      −20.       4.        3.        2.        1.        .5
10.       1.        5.       10.      100.
 .5        .6        .8         9
                     .4        .3        .2
 2.       1.        20.       5.       1.        1.        2.        2.      −2.      45.
 5   7100100000001010000000000000000000000000000010000000000000000000000000000
50.
 2.       1.        20.       5.       1.        1.        2.        2.
                                                                                  BLANK
```