DOCUMENT RESUME

ED 094 708 IR 000 903

AUTHOR	Zaremba, Joseph
TITLE	The APL/370 System Aids Instruction in Econometrics.
INSTITUTION	State Univ. of New York, Genesco.
PUB DATE	23 Oct 73
NOTE	14p.; Paper presented at the Rochester Area Colleges Workshop on the Use of Computers in Instruction (Rochester, New York, October 23, 1973)
EDRS PRICE	MF-$0.75 HC-$1.50 PLUS POSTAGE
DESCRIPTORS	*Computer Assisted Instruction; Computer Programs; *Economic Research; On Line Systems; Simulation; *Statistical Analysis; Validity
IDENTIFIERS	*Econometrics

ABSTRACT

A set of APL (A Programing Language) programs written for use on a time-sharing IBM System/370 computer and used to aid in the teaching of selected econometric concepts is described. The programs involve the simulation of a set of n observations; of a simple regression model Y = A BX U; the plotting of the Y and X simulated values; the estimation of the regression coefficients A and B and their comparison with the true population coefficients; and the plotting of the estimated and true regression lines for comparative purposes. The uses of the programs in teaching the statistical concepts and their applications to econometric problems are described, and sample outputs are shown. (WDR)

THE APL\370 SYSTEM AIDS INSTRUCTION IN ECONOMETRICS

Joseph Zaremba
Professor of Economics
State University College at Geneseo
Geneseo, New York 14454

1. Introduction. During the past year four faculty members at State University College (Geneseo), along with about 70 faculty from 15 other undergraduate institutions in New York, participated in a state-wide, 27-month, NSF supported program of experimentation and development of computer assisted instruction (CAI) using primarily APL (A Programming Language).[1] At Geneseo, this program is carried out with an IBM 2741 communications terminal with a telephone link to an IBM/370 155 computer at SUNY-Binghamton, 150 miles away. APL\370 is thus a remote-terminal, time-sharing system which uses this language with IBM System/370 computers. This paper presents some preliminary results of APL\370 in teaching econometrics.

In order to understand the APL\370 system more fully, it is useful to recall that there are two main modes of utilizing a computer: batch processing, and time-sharing processing. In batch processing, computer programs are treated as if they are in a queue waiting to be run. Instructions may be punched on cards (or some other input means such as a terminal), the numerical data that go with the instructions are then added, and the deck of cards are then placed on the card reader to wait in line for processing. In time-sharing mode, the computer inputs data from many users simultaneously (in contrast to batch processing in which one program is read in at a time), keeping up a kind of conversation with them. The users at the terminals slowly interact with the terminal, building up their programs or simply using the system's stored library of programs. Since the computer is extremely fast, the users will seldom have to wait for more than a second to obtain a response from the computer. The APL\370 system is one which provides a high degree of interaction between the user and the terminal, and is often referred to as "an interactive" approach to computer programming.

[1]APL designates the programming language that is the outgrowth of the work of K. E. Iverson, first at Harvard and then at IBM. The name comes from the initials of his book: A Programming Language (New York: Wiley, 1962)

The power and versality of APL\370 may be realized by means of a wide variety of computer programs. While we will not attempt to classify these programs, they usually can be placed under headings such as games, tutorials, simulations, and problem-solving packages. The problem-solving programs, such as The University of Alberta STATPACK2 package[2], which is available in the public libraries of most APL systems, and the New Platz Library of Statistical Programs[3], are now used routinely in teaching economic statistics at SUC (Geneseo). To use these packages the student copies the appropriate program from the library, types in the vector (or matrix) of observations and other data, and upon hitting the return key the results of the statistical analysis, which will include the estimated regression coefficients and their standard errors in a regression program, will be displayed immediately.

2. Obstacles to Teaching Econometrics. Applied econometrics consists primarily in the application of regression analysis to economic data. These data are mainly non-experimental observations--mostly in the form of economic time series--and contain well-known limitations which may make them unsuited for use by the ordinary least squares (OLS) estimating procedure.

In experimental data, for which OLS was initially intended, one attempts to obtain observations which satisfy the assumptions of regression theory. However, in the case of non-experimental data, one attempts to find a theory or model which can best utilize the existing data, or one attempts to evaluate how effectively a given regression model is able to utilize these data. Indeed, many of the problems encountered in teaching econometrics stem from attempts to utilize inadequate empirical data in regression analysis.

How can APL\370, with its rapid response and versatility, and its facility for handling arrays, aid in the improvement of teaching effectiveness in econometrics? By using the computer to simulate observations which represent real life data, and by testing various least squares estimating procedures on these data (in particular

[2]Smillie, K. W., STATPACK2: AN APL STATISTICAL PACKAGE. Department of Computer Science, The University of Alberta.

[3]Prins, J. APL: NEW PLATZ LIBRARY OF STATISTICAL PROGRAMS. 4th Ed., 1972, SUC (New Paltz).

OLS), the teaching of empirical econometrics should be greatly facilitated.

3. <u>Simulation of Economic Observations</u>. The simulations used in the programs to be discussed shortly involve stochastic processes and are called Monte Carlo simulations. An essential feature of these simulations is the generation of random numbers. Fortunately, many APL libraries contain several random number distributions and generation routines that draw random numbers according to certain probability distributions. The routines used in these programs are contained in the School of Advanced Technology (SAT) Library at SUNY-Binghamton.[4]

The advantage of APL\370 in Monte Carlo simulation is its rapid response and simplicity of use, but it should be mentioned that one disadvantage of this system is that it cannot do large simulations easily. However, for instructional purposes, this limitation is not serious.

Two programs, out of six that have been written and evaluated by a class of 16 students in econometrics, will be discussed. These programs involve the simulation of a set of n observations (sample size of n) of a simple regression model $Y = A + BX + U$; the plotting of the Y and X simulated values; the estimation of the regression coefficients A and B and their comparison with the true population coefficients; and the plotting of the estimated and true regression lines for comparative purposes. The remaining programs, not discussed here, consist of Monte Carlo simulations of various multiple regression models.

Among the objectives of these two programs are: (1) To impress the student with the fact that a sample of observations from economic time series is, in reality, a random sample from some hypothetical population; (2) That the estimated regression line is an empirical representation of the true regression line; (3) As the size of sample is increased, the estimated regression coefficients approximate more closely the true regression coefficients; (4) To emphasize the role played by the random U term in regression models; and (5) When the error term U is autocorrelated (serially

---

[4]The generation routines used in these models were written by Professor Peter Woitach of the School of Advanced Technology, SUNY-Binghamton.

correlated), the standard errors of the regression coefficients are increased relative to those of the non-serially correlated models.

Objective (1) is particularly important because textbooks in this field usually omit a worth while discussion of the underlying population from which a sample of time series observations is assumed to be drawn. Thus, students often fail to grasp the relationship between the sample that they are working with and the underlying population. However, in the case of experimental data, the student has much less difficulty in relating the sample to the population.

The first regression model to be discussed is: $Y = 300 + .3X + U$. In this model, Y represents family expenditures on consumer durables, X is family income in dollars, and U is a random error term normally distributed with mean zero and variance as specified by the user. $300 + .3X$ is the true component of Y, and U is the random component. This model reveals that as family income increases by $1, a family will spend an additional 30 cents on consumer durables, subject to the random disturbance term U.

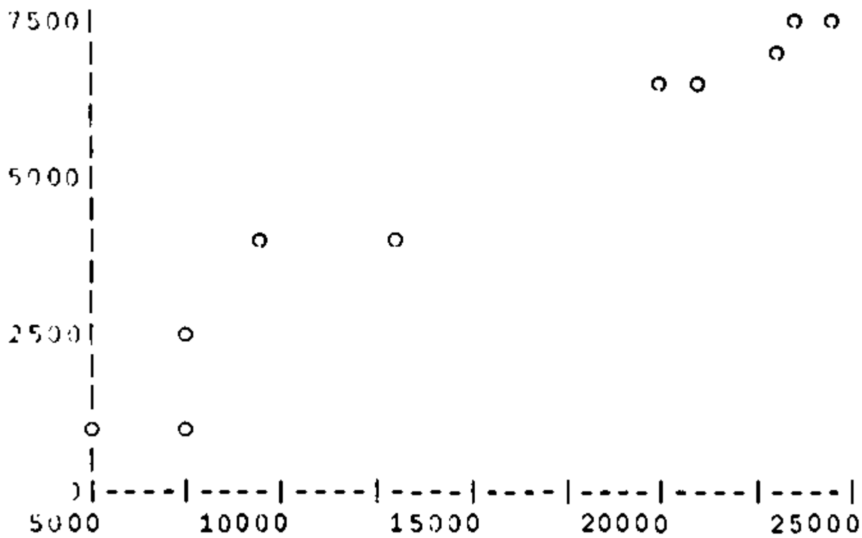
To encourage interaction between user and the computer (or terminal), a number of responses are called for, including five YES and NO responses, as the program is executed. With regard to the numerical inputs, the user is asked to select the sample size and the standard deviation of the U term. Family income is selected randomly by the computer from a uniform distribution on the interval $2000 to $25000, and the U terms are generated from a normal distribution with mean zero and standard deviation selected by the user. The computer displays these U terms, the n income values, the true component of Y, and finally the generated Y values. Simulations may be repeated with different sample sizes, income ranges, and different standard deviations.

Space limitations preclude the displaying of the APL programs in their entirety. However, portions of the print outs, beginning with the selection of sample size by the user, are displayed below for sample sizes of 10 and 30. The estimated and true regression lines plotted in the print outs should form straight lines, except for the approximation errors in the PLOT function.

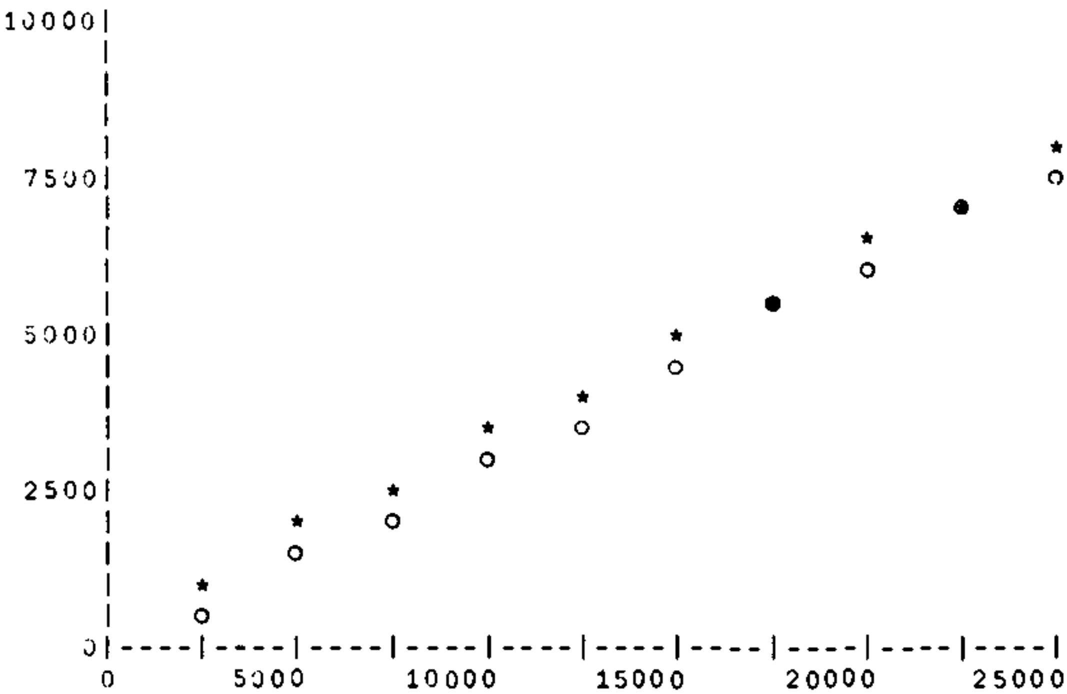
```
PLEASE SELECT A POSITIVE INTEGER BETWEEN 10 AND 30 FOR THE SIZE OF SAMPLE
⎕:
      10
THE 10 FAMILY INCOME VALUES WILL BE SELECTED RANDOMLY FROM THE INTERVAL 2000 TO 25000
       AND ARE 19903.2  5065.9  23171.5  9272.6  23691.3  24383.6  7313.9  7526.9  21032.3
      12968.7
CONSUMER DURABLE EXPENDITURES Y WILL BE GENERATED BY ASSUMING THAT THE RANDOM
       ERROR TERM U IS NORMALLY DISTRIBUTED WITH MEAN ZERO AND SPECIFIED VARIANCE.
       PLEASE SELECT A NUMBER BETWEEN 400 AND 800 FOR THE STANDARD DEVIATION OF U.
⎕:
      500
THE RANDOM ERROR TERMS U CORRESPONDING TO YOUR STANDARD DEVIATION ARE 233.3  ¯1009.8
      ¯343.15  738.65  ¯148.5  ¯184.35  ¯1302.7  ¯86.15  ¯298  ¯291.7
THE TRUE COMPONENT OF Y, NAMELY 300+0.3X, IS AS FOLLOWS 6270.96  1819.77  7251.45  3081.78
      7407.39  7615.08  2495.97  2558.07  6627.69  4190.61
EXPENDITURES ON CONSUMER DURABLES Y ARE OBTAINED BY ADDING THE TRUE COMPONENT TO
       THE RANDOM ERROR TERM U
THE ABOVE FOUR SETS OF VALUES ARE SHOWN IN THE FOLLOWING TABLE
```

X(FAMILY INCOME)	300+.3X	U	Y
19903.2	6270.96	233.3	6504.26
5065.9	1819.77	¯1009.8	809.97
23171.5	7251.45	¯343.15	6908.3
9272.6	3081.78	738.65	3820.43
23691.3	7407.39	¯148.5	7258.89
24383.6	7615.08	¯184.35	7430.73
7313.9	2495.97	¯1302.7	1193.27
7526.9	2558.07	¯86.15	2471.92
21032.3	6627.69	¯298	6329.69
12968.7	4190.61	¯291.7	3898.91

```
THE VALUES FOR Y, CONSUMER EXPENDITURES, AND X, FAMILY INCOME,
       ARE PLOTTED BELOW
```



THE ESTIMATED REGRESSION LINE, BASED ON THE SIMULATED VALUES OF Y AND X,
IS ¯304.5856596+0.3217108552X. THIS LINE IS SOMEWHAT DIFFERENT [illegible]
THE TRUE LINE 300+0.3X. BOTH LINES ARE COMPARED IN THE
FOLLOWING GRAPH WHERE THE CIRCLES REPRESENT THE
ESTIMATED LINE AND * THE TRUE LINE



IF YOU WOULD LIKE TO SIMULATE ANOTHER SET OF OBSERVATIONS WITH A
DIFFERENT SET OF INPUT VALUES, PLEASE TYPE YES. OTHERWISE TYP[illegible]

PLEASE SELECT A POSITIVE INTEGER BETWEEN 10 AND 30 FOR THE SIZE OF SAMPLE
□:

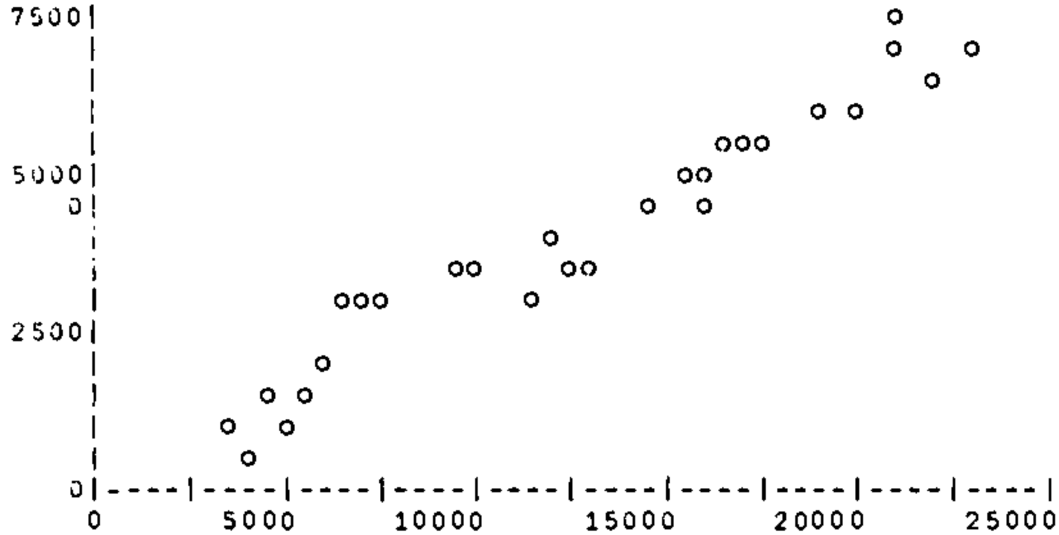
30
THE 30 FAMILY INCOME VALUES WILL BE SELECTED RANDOMLY FROM THE INTERVAL 2000 TO 25000
AND ARE 21002.6 11975.1 4796.8 15827.6 12407.5 16354.3 6731.1 16759.1
4907.2 15517.1 19187.9 7651.1 6103.2 3559.4 12934.2 17667.6 11340.3 20903.7
3955 5565 9406 14512 22010 17686 23203.7 4523.1 20075.7 9790.1 7221 16126.6
PLEASE SELECT A NUMBER BETWEEN 400 AND 800 FOR THE STANDARD DEVIATION OF U.

□:

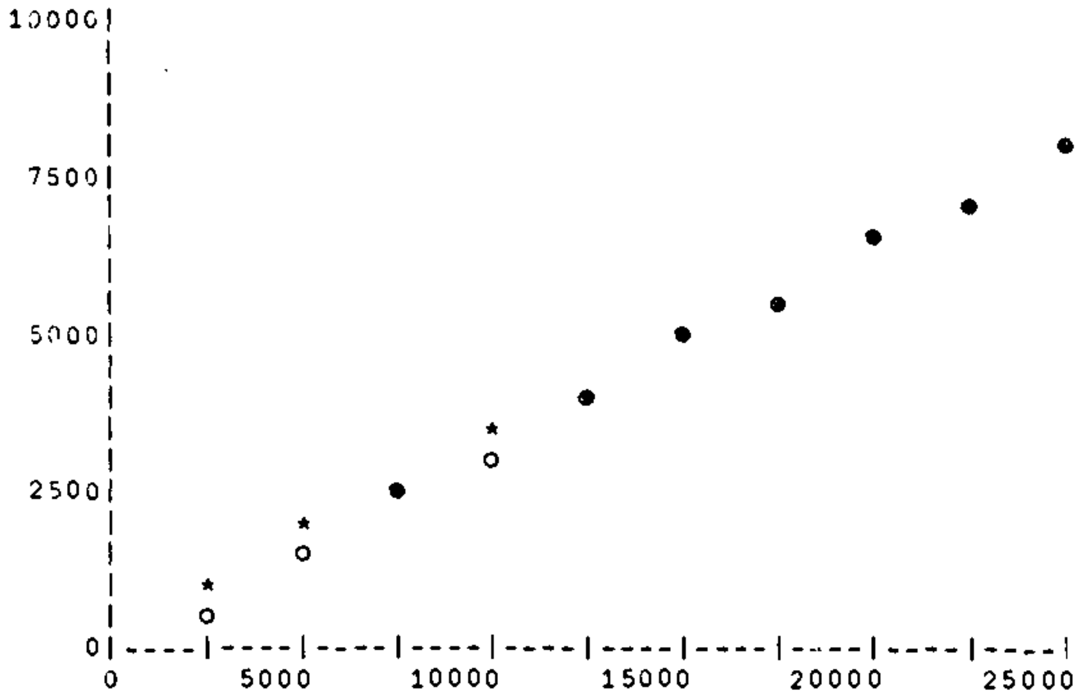
500
THE TRUE COMPONENT OF Y, NAMELY 300+0.3X, IS AS FOLLOWS 6600.78 3892.53 1739.04 5048.28
4022.25 5206.29 2319.33 5327.73 1772.16 4955.13 6056.37 2595.33 2130.96
1367.82 4180.26 5600.28 3702.09 6571.11 1486.5 1969.5 3121.8 4653.6 6903
5605.8 7261.11 1656.93 6322.71 3237.03 2466.3 5137.98
CONSUMER EXPENDITURES Y ARE OBTAINED BY ADDING THE TRUE COMPONENT TO THE RANDOM TERM U.
THE ABOVE FOUR SETS OF VALUES ARE SHOWN IN THE FOLLOWING TABLE

X(FAMILY INCOME)	300+.3X	U	Y
21002.6	6600.78	725.2	7325.98
11975.1	3892.53	345	4237.53
4796.8	1739.04	-710.85	1028.19
15827.6	5048.28	-729.45	4318.83
12407.5	4022.25	-612.35	3409.9
16354.3	5206.29	375.85	5582.14
6731.1	2319.33	890.1	3209.43
16759.1	5327.73	130.15	5457.88
4907.2	1772.16	-654.1	1118.06
15517.1	4955.13	259.75	5214.88
19187.9	6056.37	37.05	6093.42
7651.1	2595.33	189.2	2784.53
6103.2	2130.96	-375.05	1755.91
3559.4	1367.82	-538.2	829.62
12934.2	4180.26	-580.15	3600.11
17667.6	5600.28	86.7	5686.98
11340.3	3702.09	-477.8	3224.29
20903.7	6571.11	394.75	6965.86
3955	1486.5	-839.15	647.35
5565	1969.5	-714.15	1255.35
9406	3121.8	246.05	3367.85
14512	4653.6	27.65	4681.25
22010	6303	-170.5	6732.5
17686	5605.8	-60.1	5545.7
23203.7	7261.11	-14.1	7247.01
4523.1	1656.93	71.1	1728.03
200[illegible]	6322.71	-[illegible].5	[illegible]
[illegible]	[illegible]	[illegible]	[illegible]
[illegible]	2466.3	587.05	3053.35
[illegible]	5137.98	-854.2	4283.7[illegible]

THE [illegible] ARE PLOTTED BELOW.



THE ESTIMATED REGRESSION LINE, BASED ON THE SIMULATED VALUES OF Y AND [illegible]
IS -77.83664007+0.322960493X. THIS LINE IS SOMEWHAT DIFFERENT [illegible]
THE TRUE LINE 300+0.3X. BOTH LINES ARE COMPARED IN THE
FOLLOWING GRAPH WHERE THE CIRCLES REPRESENT THE
*ESTIMATED LINE AND * THE TRUE LINE*



IF YOU WOULD LIKE TO SIMULATE ANOTHER SET OF OBSERVATIONS WITH A
DIFFERENT SET OF INPUT VALUES, PLEASE TYPE YES. OTHERWISE TY[illegible]

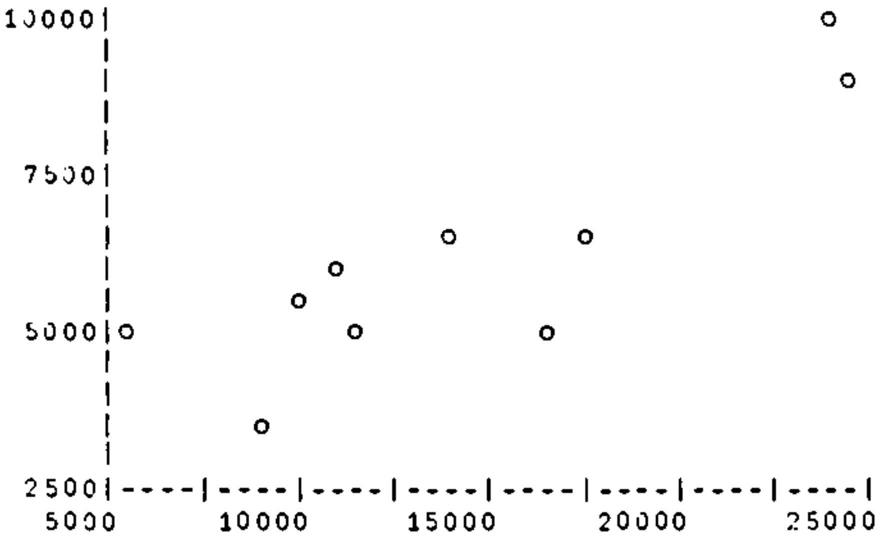
The second regression model is $Y = 300 + .3X + U$ where U is now serially correlated according to the autocorrelation scheme $U(t) = R x U(t-1) + E(t)$. U(t) is the error term at time t, U(t-1) is the error term at time t-1, R is a constant with absolute value less than or equal to 1, and E(t) is an error term normally distributed with mean zero and specified variance. This formula generates the last n-1 U terms based on an initial value for U, called U(1), which is selected randomly from the interval (-250,250). The Y and X variables are identical to those in the first model. The students are asked to type in the standard deviation of E, the value of R, the upper and lower limits of family income, and the size of sample.

Many real life relationships between two variables correspond to our second model and the simulation of observations based on this model should enable the student to gain valuable insights regarding the proper use of empirical data. In reality, the U term contains the effects of all variables on the dependent variable Y except the influence of the independent variable X. Because in the real world a great number of factors influence the dependent variable, the U term in a regression model could be very important.

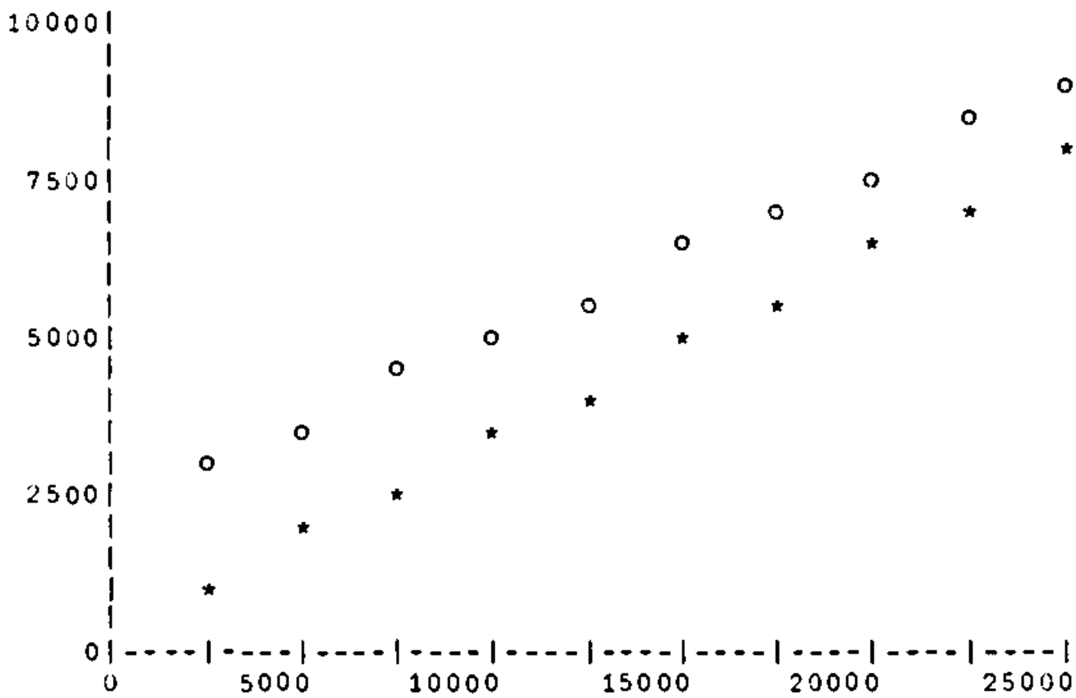
It can be shown that when the error term U is autocorrelated by the above scheme, the estimates of the regression coefficients by OLS are unbiased, but the standard errors of these coefficients will be very large relative to non-autocorrelated U terms. Hence, hypothesis testing and interval estimation concerning the regression coefficients may not be reliable.

Simulations of Y (family expenditures) and X (family income) observations according to our second model are shown below in print outs of programs based on this model. The observations and plottings are displayed in much the same manner as those of the first print out. For both sample sizes of 10 and 30, the value of R is 1, the standard deviation of E is 500, and the lower and upper limits of family income are 2000 and 25000, respectively.

```
IN ORDER TO GENERATE A SET OF N OBSERVATIONS WITH AUTOCORRELATED U TERMS OF THE
      FORM U(T)=R×U(T-1)+E(T) YOU WILL BE ASKED TO SUPPLY CERTAIN INFORMATION.  PLEASE
      SELECT AN INTEGER BETWEEN 10 AND 30 FOR THE SAMPLE SIZE
⎕:
     10
WHAT IS THE LOWER LIMIT OF THE FAMILY INCOME RANGE?
⎕:
     2000
PLEASE SELECT A NUMBER FOR THE UPPER LIMIT OF FAMILY INCOME
⎕:
     25000
THE 10 INCOME VALUES SELECTED AT RANDOM FROM THE INTERVAL 25000-2000 ARE 16257.7  8888.5
     17361.7  24434.2  11252.9  10928.6  14134.8  9958  24190.4  5512.1
THE ERROR TERMS U ARE DETERMINED BY THE FORMULA U(T)=R×U(T-1)+E(T).  THE INITIAL
      VALUE FOR U, DENOTED U(1) AND SELECTED FROM THE INTERVAL -250 TO 250, IS ¯47
PLEASE SELECT A NUMBER BETWEEN 200 AND 700 FOR THE STANDARD DEVIATION OF E
⎕:
     500
PLEASE SELECT A NUMBER BETWEEN ¯1 AND 1 FOR THE VALUE OF R
⎕:
     1
THE 10 AUTOCORRELATED U TERMS ARE ¯47  343  994.7  1394.1  1467.9  2254.7  1952.6  2347.4
     2255.65  2999.25
THE SIMULATED EXPENDITURES ON CONSUMER DURABLES Y ARE FOUND BY ADDING 300+.3X
      TO THE U TERMS AND ARE AS FOLLOWS 5130.31  3309.55  6503.21  9024.36  5143.77
     5833.28  6493.04  5634.8  9812.77  4952.88 THESE VALUES ARE PLOTTED AGAINST
      FAMILY INCOME ON THE FOLLOWING PAGE
```



THE ESTIMATED REGRESSION LINE IS Y=2344.418432+0.2686403666X, AS COMPARED WITH THE TRUE LINE 300+0.3X. THE ESTIMATED LINE, DENOTED BY o, AND THE TRUE LINE, DENOTED BY *, ARE PLOTTED BELOW



IF YOU WOULD LIKE TO SIMULATE ANOTHER SET OF OBSERVATIONS WITH A DIFFERENT SET OF INPUT VALUES, PLEASE TYPE YES. OTHERWISE, TYPE NO

IN ORDER TO GENERATE A SET OF N OBSERVATIONS WITH AUTOCORRELATED U TERMS OF THE
FORM U(T)=R×U(T-1)+E(T) YOU WILL BE ASKED TO SUPPLY CERTAIN INFORMATION. PLEASE
SELECT AN INTEGER BETWEEN 10 AND 30 FOR THE SAMPLE SIZE

⎕:

30

WHAT IS THE LOWER LIMIT OF THE FAMILY INCOME RANGE?

⎕:

2000

PLEASE SELECT A NUMBER FOR THE UPPER LIMIT OF FAMILY INCOME

⎕:

25000

THE 30 INCOME VALUES SELECTED AT RANDOM FROM THE INTERVAL 25000-2000 ARE 9856.8 2280.6
3159.2 20048.1 22230.8 12538.6 20117.1 20751.9 23833.9 7338.3 8844.8 12872.1
3350.1 21248.7 19990.6 18916.5 13881.8 9164.5 20634.6 10588.2 18714.1
19912.4 20257.4 20110.2 16370.4 22290.6 23638.4 13628.8 18424.3 11839.4
THE ERROR TERMS U ARE DETERMINED BY THE FORMULA U(T)=R×U(T-1)+E(T). THE INITIAL
VALUE FOR U, DENOTED U(1) AND SELECTED FROM THE INTERVAL -250 TO 250, IS ¯92.625
PLEASE SELECT A NUMBER BETWEEN 200 AND 700 FOR THE STANDARD DEVIATION OF E

⎕:

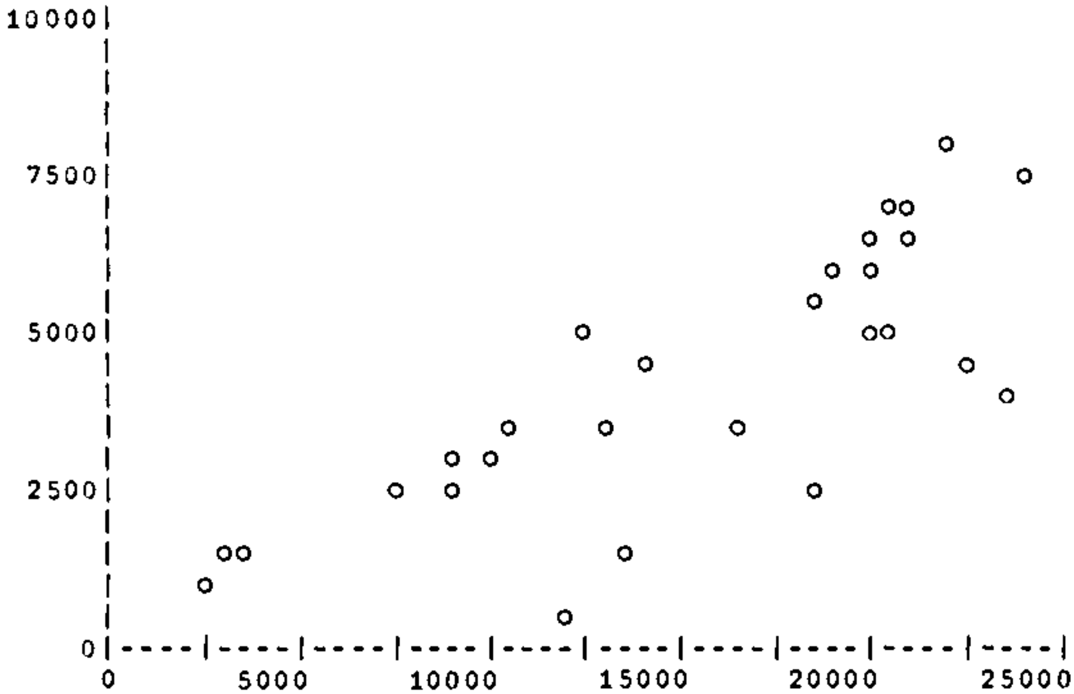
500

PLEASE SELECT A NUMBER BETWEEN ¯1 AND 1 FOR THE VALUE OF R

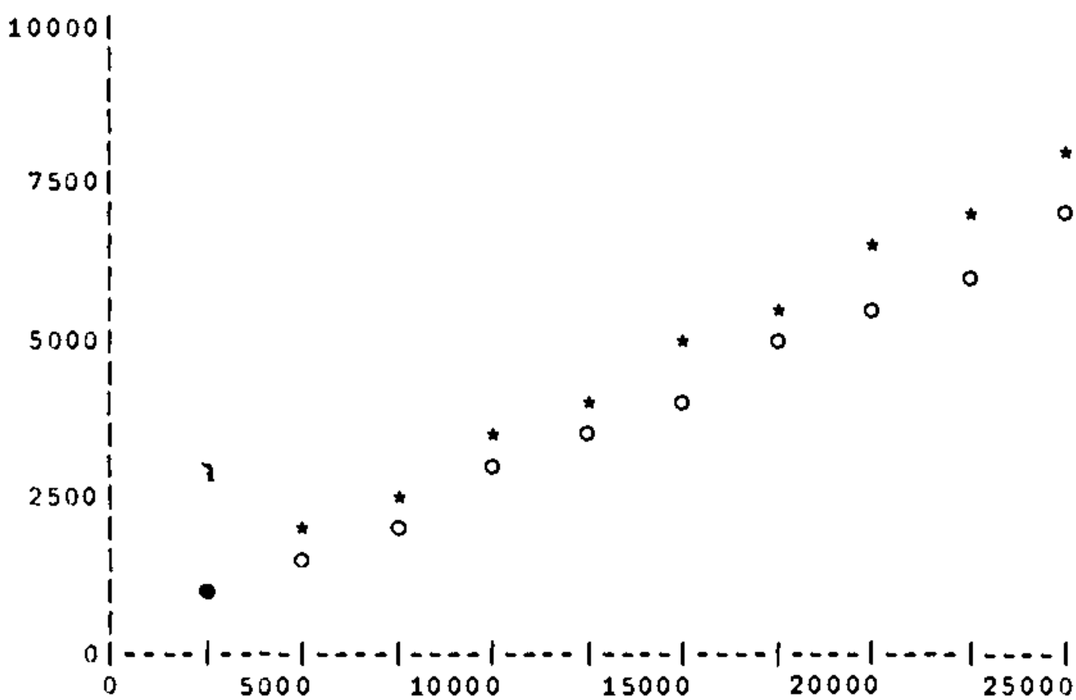
⎕:

1

THE 30 AUTOCORRELATED U TERMS ARE ¯92.625 ¯191.425 66.175 405.725 996.925 894.275
¯190.225 24.725 50.225 4.925 ¯338.225 ¯478.525 24.125 239.725 29.325 19.625
¯20.175 ¯112.025 704.275 153.925 ¯410.025 ¯1064.375 ¯1412.925 ¯1204.675
¯1958.975 ¯2429.775 ¯3143.625 ¯2957.675 ¯3373.625 ¯3339.725
THE SIMULATED EXPENDITURES ON CONSUMER DURABLES Y ARE FOUND BY ADDING 300+.3X
TO THE U TERMS AND ARE AS FOLLOWS 3164.415 792.755 1313.935 6720.155 7966.165
4955.855 6144.905 6550.295 7500.395 2506.415 2615.215 3683.105 1329.155
6914.335 6326.505 5994.575 4444.365 2937.325 7194.655 3630.385 5504.205
5208.745 4964.295 5128.385 3252.145 4557.405 4247.895 1430.965 2453.665
512.095 THESE VALUES ARE PLOTTED AGAINST
FAMILY INCOME ON THE FOLLOWING PAGE



*THE ESTIMATED REGRESSION LINE IS Y=145.9801195+0.268972529X, AS COMPARED WITH THE TRUE LINE 300+0.3X. THE ESTIMATED LINE, DENOTED BY O, AND THE TRUE LINE, DENOTED BY *, ARE PLOTTED BELOW*



IF YOU WOULD LIKE TO SIMULATE ANOTHER SET OF OBSERVATIONS WITH A DIFFERENT SET OF INPUT VALUES, PLEASE TYPE YES. OTHERWISE, TYPE NO

4. <u>Analysis of Results</u>. With regard to the simulations of model 1 (pp. 5-8), it can be seen that as the sample size increased from 10 to 30, the estimated and true regression lines converged. In general, the students also experienced this same tendency when they executed the program.

Even more revealing is the comparison of the first plottings of both print outs (pp. 6 and 11), and also the second plottings in these print outs (pp. 8 and 13). The first comparison is between the simulated values of Y and X for each model for sample sizes of 10, and the second is a comparison between the simulated Y and X values for sample sizes of 30. It will be noted that the dispersion of the Y observations generated from the autocorrelated model is much wider than that from the non-autocorrelated model--a result that can be expected on the basis of statistical theory. A comparison of the two plottings in the second print out (pp. 11 and 13) also reveals that as the sample size increases, the dispersion of the Y simulated values also increases.

All 16 students in undergraduate econometrics in the spring of 1973 executed these programs enthusiastically. Everyone believed that they learned something about econometrics that they failed to learn from the text and lectures and they expressed a desire to see greater use made of this technique for instructional purposes. The objectives of these two programs, outlined earlier, were largely achieved.

The students mentioned that they learned the following specific things from running these programs: (1) The realization that the estimated line does not coincide with the true regression line; (2) As the size of sample increased, these two lines approached one another; (3) A greater appreciation of the role played by the U term in regression analysis; (4) The nature of randomness, as illustrated by the display of the simulated observations, both graphically and tabularly; and (5) The wide dispersion of the dependent variable as autocorrelation is introduced into regression models.

The main criticisms of the programs were that they are somewhat lengthy, and that not enough input was required by the user.