NUREG/CR-2350
SAND81-1978

Sensitivity-Analysis Techniques: Self-Teaching Curriculum

NUREG/CR--2350

DE82 017558

Manuscript Completed: January 1982
Date Published: June 1982

Prepared by
R. L. Iman, W. J. Conover*

Sandia National Laboratories
Albuquerque, NM 87185

*Texas Tech University

Prepared for
Division of Waste Management
Office of Nuclear Material Safety and Safeguards
U.S. Nuclear Regulatory Commission
Washington, D.C. 20555
NRC FIN A1158

DISCLAIMER

ABSTRACT

This self teaching curriculum on sensitivity analysis techniques consists of three parts:

1) Use of the Latin Hypercube Sampling Program [Iman, Davenport and Ziegler, Latin Hypercube Sampling (Program User's Guide), SAND79-1473, January 1980].

2) Use of the Stepwise Regression Program [Iman, et al., Stepwise Regression with PRESS and Rank Regression (Program User's Guide) SAND79-1472, January 1980].

3) Application of the procedures to sensitivity and uncertainty analyses of the groundwater transport model NWFT/DVM [Campbell, Iman and Reeves, Risk Methodology for Geologic Disposal of Radioactive Waste - Transport Model Sensitivity Analysis; SAND80-0644, NUREG/CR-1377, June 1980: Campbell, Longsine, and Reeves, The Distributed Velocity Method of Solving the Convective-Dispersion Equation, SAND80-0717, NUREG/CR-1376, July 1980].

This curriculum is one in a series developed by Sandia National Laboratories for transfer of the capability to use the technology developed under the NRC funded High Level Waste Methodology Development Program (NRC FIN. No. A-1192). The technology transfer process is carried out under NRC Fin No. A-1158.

TABLE OF CONTENTS

TABLE OF CONTENTS (cont'd)

TABLE OF CONTENTS (cont'd)

TABLE OF CONTENTS (cont'd)

TABLE OF CONTENTS (cont'd)

LIST OF FIGURES

PART ONE:

LIST OF FIGURES (Continued)

LIST OF FIGURES (Continued)

LIST OF FIGURES (Continued)

LIST OF FIGURES (Continued)

LIST OF TABLES

PART ONE

TUTORIAL ON THE LATIN HYPERCUBE SAMPLING PROGRAM

The Purpose of the Course

The purpose of this tutorial is to demonstrate how to draw multivariate random samples, using either Random Sampling or using Latin Hypercube Sampling, where the multivariate random sample may have any specified marginal distributions and any specified correlation matrix. This tutorial shows how to obtain such a sample manually, and how to use the Latin hypercube computer program to accomplish the same task. A comparison between Random Sampling and Latin Hypercube Sampling is made to show some of the relative benefits of using a Latin Hypercube Sample.

The Importance of Good Sampling Techniques

Some physical processes are difficult to study directly, for various reasons, and are therefore observed indirectly through the use of mathematical models. The mathematical models are often so complex that they are amenable to solution only through the use of numerical methods on a computer. Even then the solution may involve a considerable amount of computer time, so care is needed in selecting the input variables in such a way that the most important information is obtained concerning the output variable. This suggests the use of efficient statistical methods for the design and analysis of pseudo-data generated through the use of such models. Some of those methods are described in detail in the following sections.

The Nuclear Waste Repository Model

Consider the model for simulating geologic conditions in a repository for nuclear waste. The input variables may be random variables or may be parameters whose values are unknown but may be known to lie in given intervals with specified probabilities. In either case, the input variables are subject to uncertainties that may be described by means of a probability distribution. In Figure 1 a diagram of a hypothetical nuclear waste repository is given, and four input variables are shown along with their hypothesized probability distributions. Each probability distribution is specified by name and by the lower limit $\underline{a}$ and upper limit $\underline{b}$. In the normal distribution $\underline{a}$ and $\underline{b}$ represent truncation of the usual normal distribution at the $\overline{.001}$ and $\overline{.999}$ quantiles. The upper limit for the lognormal and loguniform distributions are taken to be the .999 quantiles. The convenience of working with a finite range for each variable considerably outweighs the disadvantage of working with a slightly truncated form of a standard distribution.

Several Types of Input Distributions

The normal distribution and the uniform distribution are well known distributions which are symmetric. Typical density functions for these two distributions are given in Figure 1. The lognormal distribution is a unimodal distribution, skewed to the right, which is often used to represent random variables that assume only positive values. The logarithm of a lognormally distributed random variable is a random variable with a normal distribution. A loguniform distribution has many of the same properties as a lognormal distribution, such as being unimodal, skewed to the right, and nonnegative. The shape of the distribution is slightly different, with the right tail of the distribution being considerably heavier than the right tail of a lognormal distribution. If a random variable has a loguniform distribution, its logarithm has a uniform distribution. Instead of these four distributions, any probability distribution may be specified for the input variables. These distributions are simply the ones used most frequently in this type of model.

The Importance of Accurate Input Distributions

In any model such as the waste isolation model the output from the model is the item of interest. There are uncertainties attached to the output because there are uncertainties inherent in the input. Thus the output is expressible only in terms of its probability distribution, or properties such as the mean, standard deviation, median, quartiles or other quantities. An accurate representation of the output requires an accurate representation of the input. Therefore good answers to questions concerning the output require accurate representations of the input distributions. Since the quantities represented by the input variables may possess some correlation in nature, that same correlation should be reflected in the selection of input variables for the model. The requirements imposed thus far, i.e., specified distributions on each variable and specified correlations between variables, require some nonstandard statistical methodology. The methods presented in this tutorial have been developed specifically for models such as this one, so they will probably not be familiar to the reader.

The Simplified Black Box Model

In order to present the statistical methodology in a clear, uncluttered manner, details which are not relevant to the statistical methodology are suppressed in this tutorial. One such detail is the model itself. Although the model is the most important link in the study of a physical system, the proper development and verification of the model is the responsibility of geologists, physicists, engineers and other experts. From the statistician's viewpoint the model is viewed as a "black box," with many input variables and one or more output variables. Figure 2 shows a diagram of the model from a statistician's viewpoint,

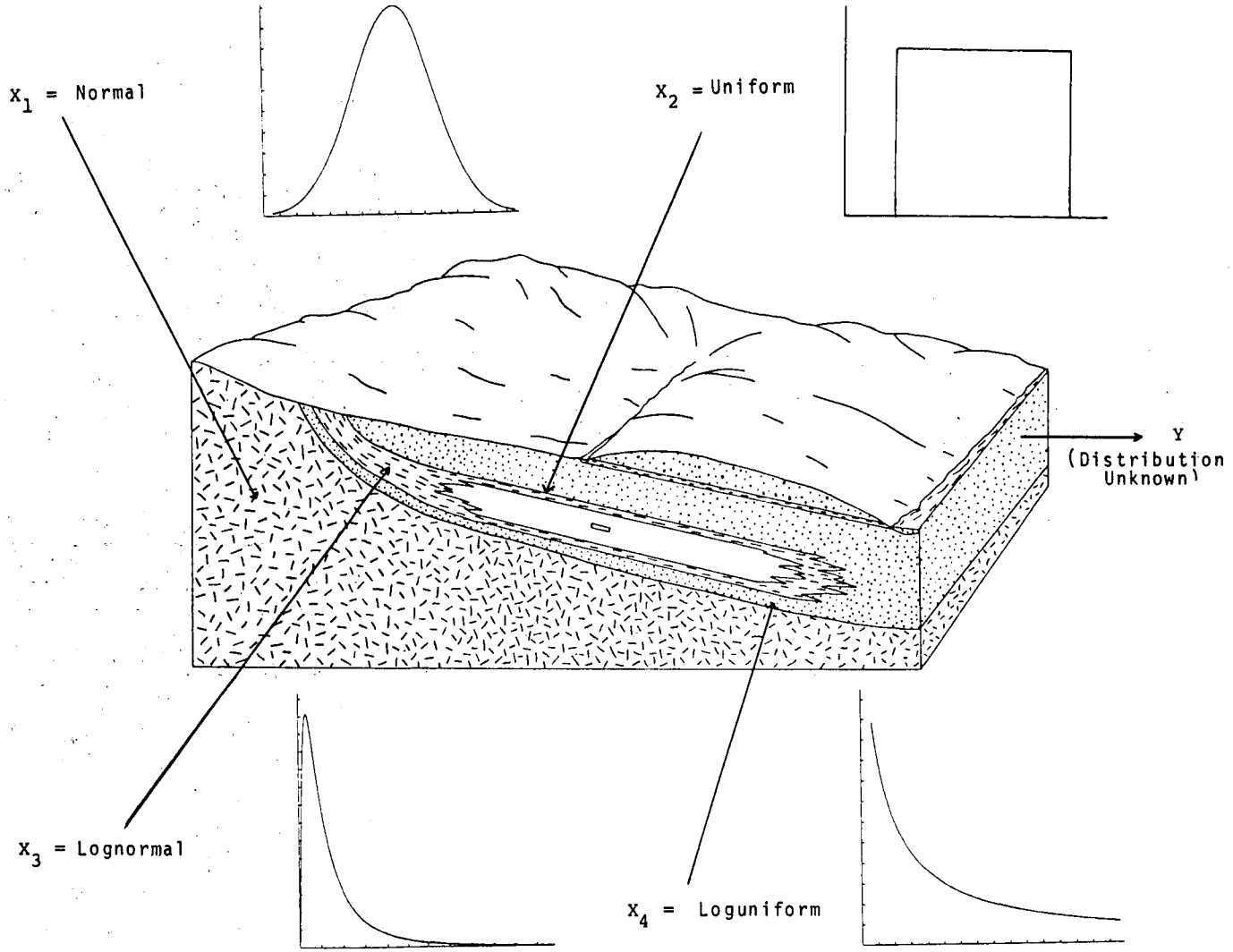


Figure 1. Some Input Variables for a Hypothetical Nuclear Waste Repository.

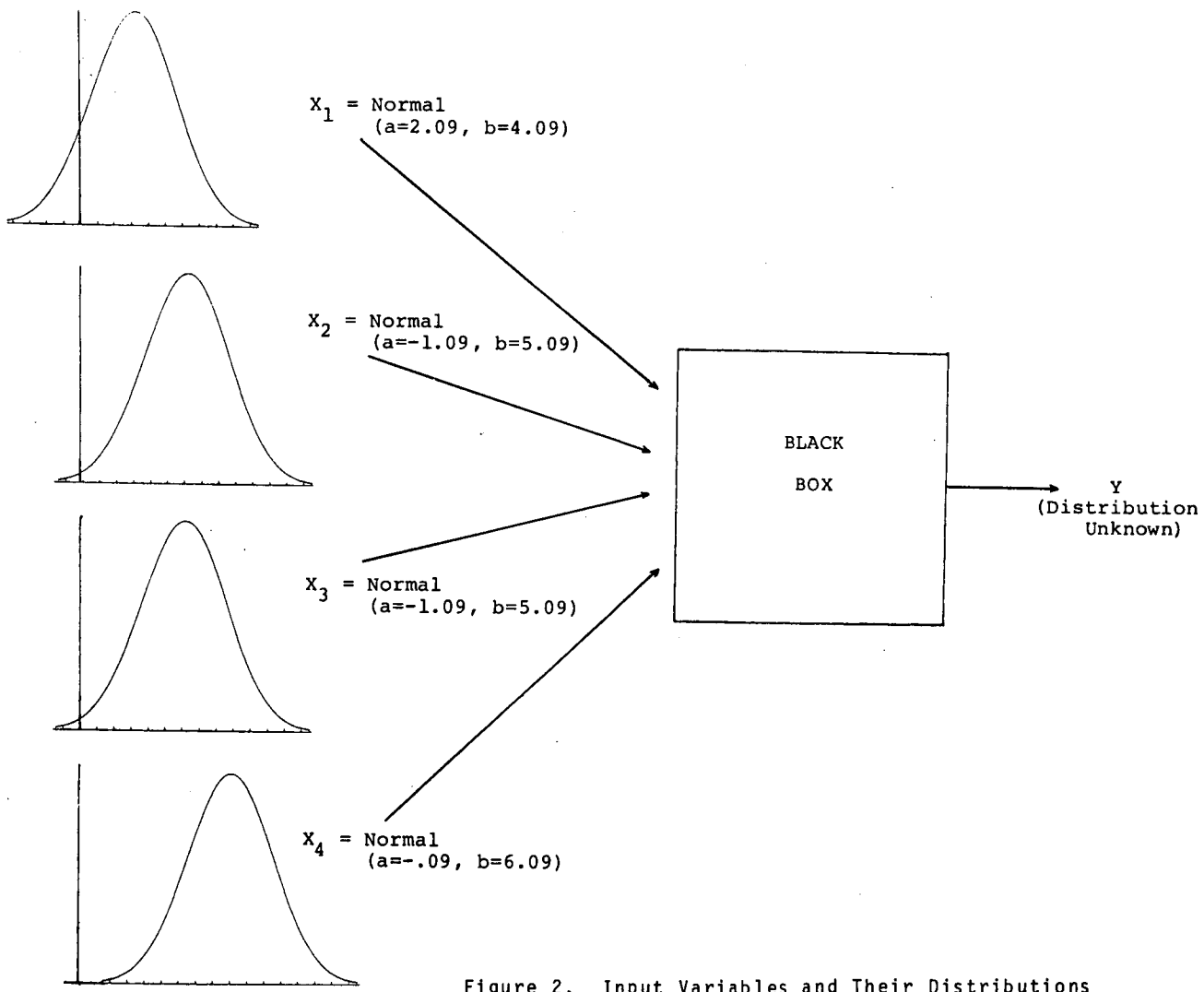


Figure 2. Input Variables and Their Distributions for a Hypothetical Black Box Model.

where there are four input variables and where the input variables have specified probability distributions, all normal in this case. The statistician's job is to understand the desired objectives of the study, and to use (and/or develop) methods for achieving those objectives. Some of the desired objectives of this study are given as follows.

Objectives of a Simulation Study

To assess the probability of the output exceeding specified limits.

For some values of the input variables the output of a model may exceed the limits of acceptability as imposed by regulatory agencies. How likely is this to happen? Because the input variables have uncertainties associated with them, the exceedance probability can only be estimated on the basis of several runs in which input variables are selected and the output variable is observed. A method for obtaining a confidence interval on this probability is also needed. If several valid methods are available for finding such a confidence interval, the method that gives the smallest interval is obviously the best method to use for achieving this objective.

To determine the sensitivity of the output to the various input variables.

If some input variables are very influential on the output variable, those input variables require close study in any actual site selection decisions. Assumptions regarding the distributions of those variables also need careful consideration. On the other hand any input variables that show little or no influence on the output variable are not very important to study from a cost effective standpoint, and assumptions regarding their distributions are not as critical. Statistical methods are needed for measuring the relative importance of the input variables on the basis of several runs of the model. Note that some variables may be important at some time points but not important at other time points, so the method of selecting input values for the various runs of the model should be flexible enough to allow this determination. The elimination of nonsignificant input variables may result in a substantial simplification of the model, and a sharper focus on the more relevant aspects of the model.

On the Latin Hypercube Sampling Program

The Latin Hypercube Sampling Program was written to enable a researcher to select input variables according to any of several different methods. It is necessary for the user of this program to specify several items, including the input distributions, the correlation matrix of the input variables, and the type of sampling procedure desired. The program then takes care of obtaining numerical quantities to use as input variables for the model, where those numerical quantities resemble

values of random variables with the specified probability distributions, with a correlation structure as specified by the user, and selected according to the specified sampling scheme. As the name of the program suggests, one of the options for sampling is Latin Hypercube Sampling, which is a very useful sampling scheme developed specifically for problems such as this one. However, the user may specify Random Sampling instead, which is a frequently used sampling procedure. Variations of these sampling procedures are available as options in this program, however, attention will be focused primarily on these two options.

Options for Input Distributions

The input distribution is specified separately for each input variable. There are five options available for input distributions; normal, uniform, lognormal, loguniform, and a user-specified distribution. The first four distributions are built into the program and are very easy to use. These were discussed earlier. The user-specified distribution requires that a subroutine be written to supply distributions other than those four.

The Distribution Function vs. the Density Function

Although it is more usual in statistics to think in terms of density functions when describing the distribution of a random variable, there are definite advantages of considering distribution functions (CDFs) in this tutorial. A CDF represents the cumulative probability associated with a random variable. That is, if f(x) is the density function of a continuous random variable X, then the distribution function F(x) represents the cumulative probability up to the value x,

$$F(x) = P(X \leq x) = \int_{-\infty}^{x} f(t)\, dt$$

While a random variable must be continuous in order to possess a density function, all random variables possess distribution functions, whether they are continuous, discrete, or some combination of continuous and discrete. Figures 3, 4, 5 and 6 illustrate distribution functions and density functions for particular normal, uniform, lognormal, and loguniform distributions. Figure 7 presents the estimated distribution function of the number of boreholes present in a randomly selected 1100 acre tract which is underlain by bedded salt and which has at least one borehole present. Note that this is a discrete distribution function.

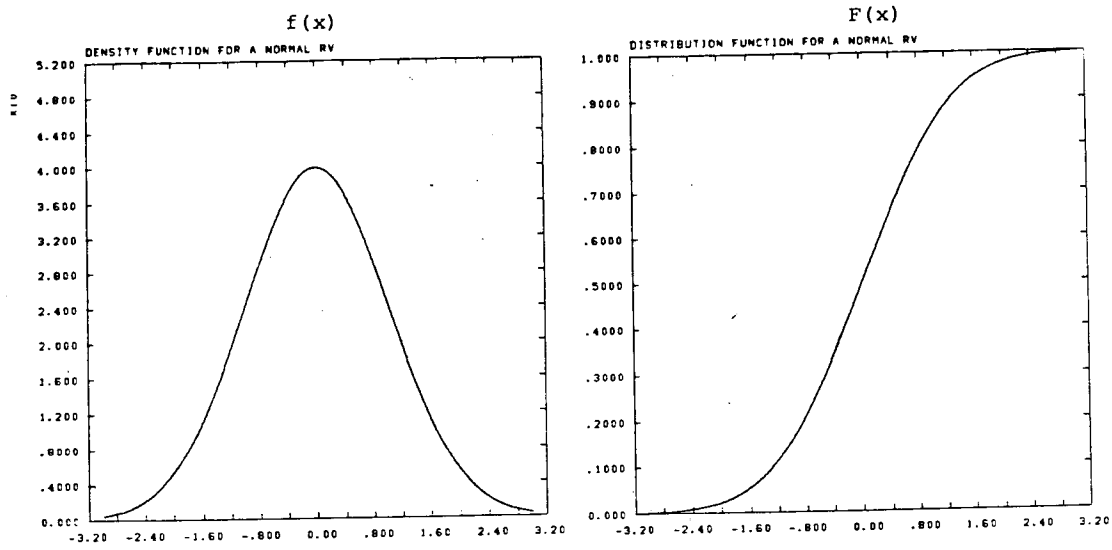


Figure 3. A Normal Density Function f(x) and its Corresponding Distribution Function F(x).

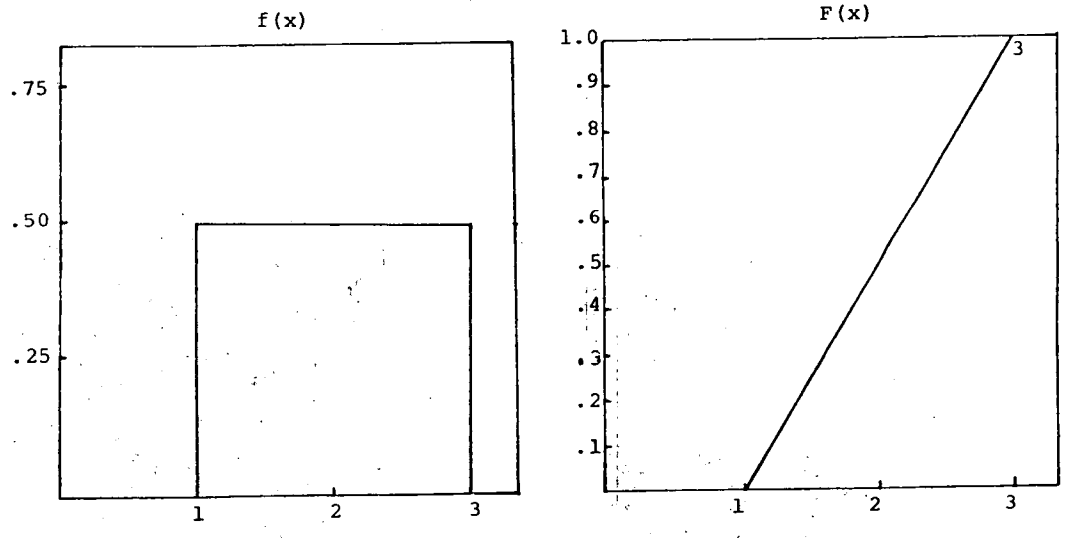


Figure 4. A Uniform Density Function f(x) and its Corresponding Distribution Function F(x).

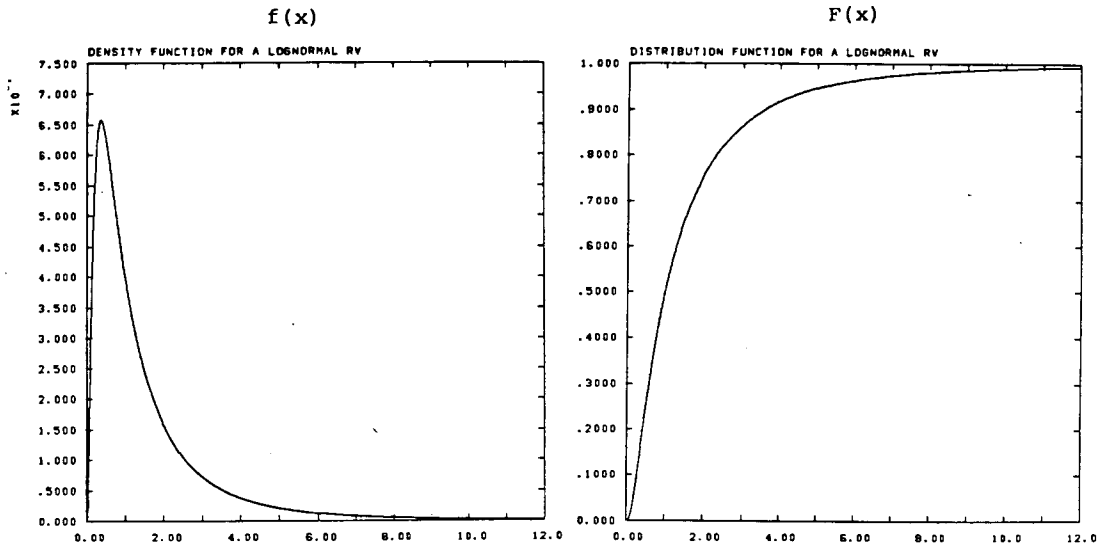


Figure 5. A Lognormal Density Function f(x) and its Corresponding Distribution Function F(x).

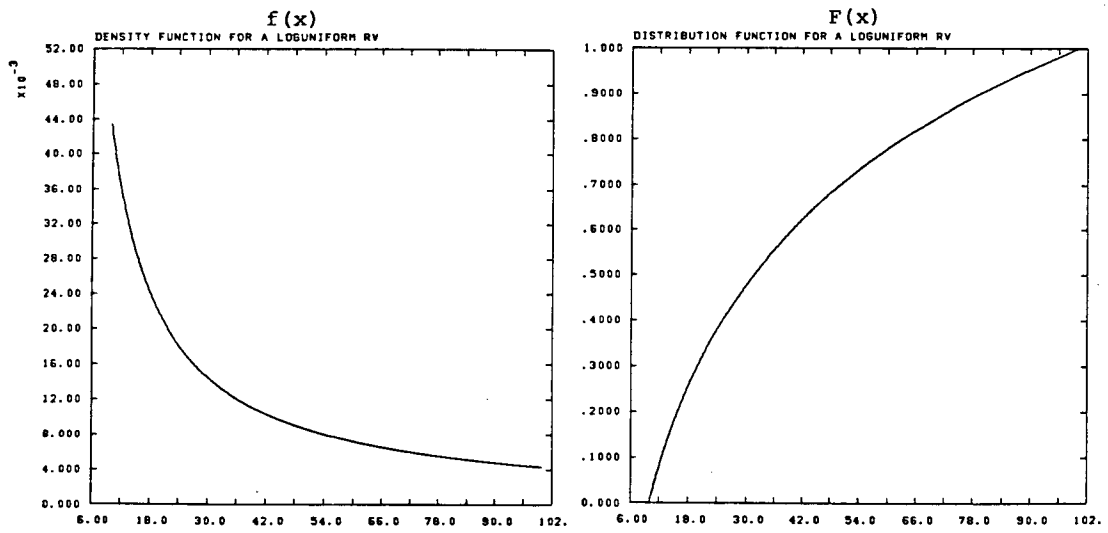


Figure 6. A Loguniform Density Function f(x) and its Corresponding Distribution Function F(x).

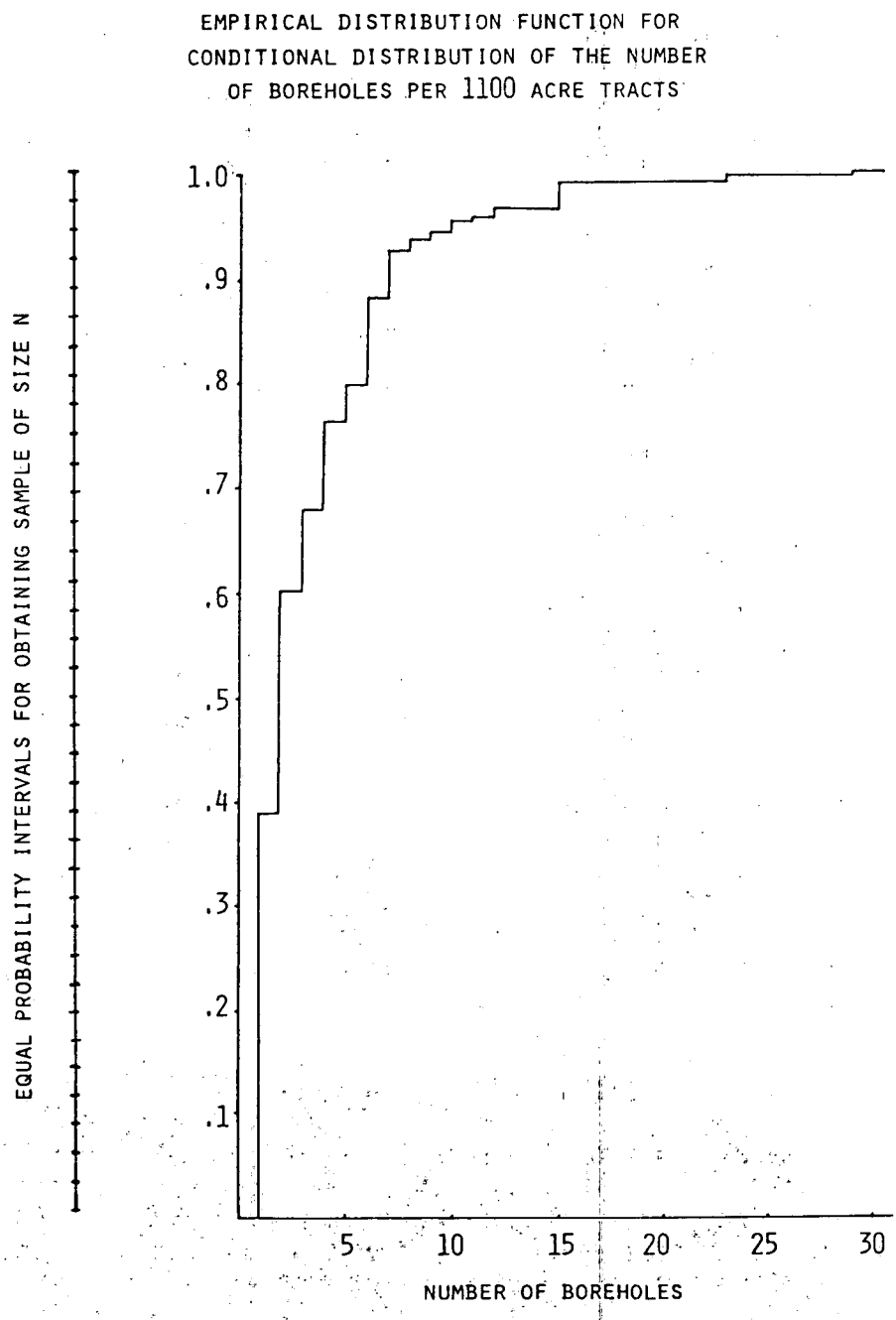


Figure 7. A Distribution Function Estimated From a Large Number of Sample Observations.

Drawing a Random Sample (Illustration)

A hypothetical distribution function is given in Figure 8 for the purposes of illustrating the principles behind drawing a random sample. If a random sample of size N = 2 is to be drawn from F(x), the following steps are followed.

1. Draw N = 2 random uniform values. A table of random digits such as Table 1 may be used for this purpose. Enter the table at some randomly selected row (suggestion: let the row number equal the last two digits of your social security number) and a random column (perhaps the third digit from the end of your social security number) and record two sets of numbers, two digits each, reading across. If row 31, column 3 is selected, the numbers are 87 and 91.

2. Convert these to numbers between 0 and 1. The simplest way to do this is to divide by 100, which converts 87 to .87 and 91 to .91.

3. Use Figure 8 to find F^{-1} for these numbers. The numbers .87 and .91 are found on the vertical axis in Figure 8, and the inverse function F^{-1} of F(x) is used to convert these to 2.13 and 2.34. These two numbers 2.13 and 2.34 are the random sample of size 2 from the probability distribution given by F(x).

Note that by chance both of the numbers in the random sample in the example happened to be near the upper end of the range of possible values of the random variable being sampled, because the two numbers from Table 1 happened to be close to 100. This is the nature of random samples; no guarantee is given that the numbers in the sample will be spread out over the range of possible values of the random variable. For this reason, the following method of sampling, called stratified sampling, is often preferred.

Drawing a Stratified Sample (Illustration)

A hypothetical distribution function F(x) (the same one used in Figure 8) is given in Figure 9 to illustrate the principle behind stratified sampling. For a stratified sample of size N = 2, one observation is sampled at random from the lower half (in a probability sense) of the distribution and one value is sampled at random from the upper half.

1. Draw N = 2 random uniform values. For simplicity the same numbers 87 and 91 will be used again.

2. Convert the first number to a number between 0 and .5 and the second to a number between .5 and 1.0. First, each number is converted to a number between 0 and .5 by dividing by 200. Then .5 is added to the second number.

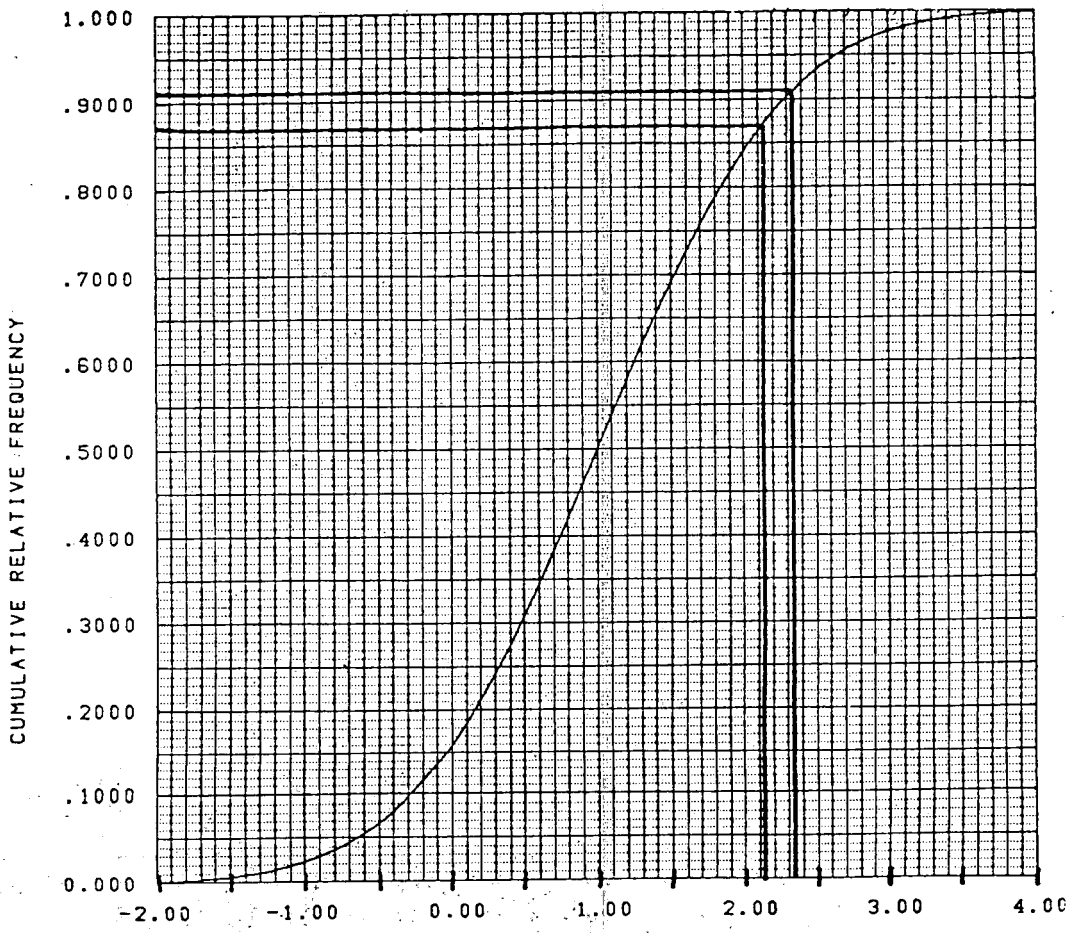


Figure 8. A Hypothetical Distribution Function F(x) for Use in Drawing a Random Sample of Size N =2.

$$\frac{87}{200} = .435$$

$$\frac{91}{200} + .5 = .455 + .5 = .955$$

3. Use Figure 9 to find F^{-1} for these numbers. The numbers .435 and .955 are found on the vertical scale in Figure 9, and the inverse function F^{-1} of F(x) is used to convert these to a stratified sample of size 2 from F(x). The sample consists of the two numbers 0.84 and 2.70. Note that the first number is in the lower half of the distribution and the second number is in the upper half of the distribution, in a probability sense.

For a stratified sample of size N, the vertical axis of Figure 9 would be divided into N equal intervals between 0 and 1, one observation would be sampled at random from each interval using uniform random numbers, and these would be converted to a random sample from F(x) through the use of an inverse function.

More Exact Normal Values from Table 2

The graphical method for finding x, given F(x), that was illustrated in the preceding examples is simple and straightforward. It may be used with the graph of any distribution function. The only limitation of such a method is that graphical methods are good to 2 or 3 decimal place accuracy at best. Of course, if the distribution function is discrete, as in Figure 7, whole integer accuracy may be sufficient.

The distribution function used in the illustrations happens to be normal with mean 1 and variance 1. Therefore, the exact values for F^{-1}, column (d) in Figures 10 and 12, may be found from Table 2. Round the value in column (c) to three decimal places. Enter Table 2 to get Φ^{-1}, the inverse for a standard normal distribution ($\mu = 0$, $\sigma = 1$). Then add 1 to get the inverse for F(x), because $\mu = 1$. In general, multiplication by σ and addition of μ, in that order, converts from Φ^{-1} to any normal random variable with mean μ and variance σ^2.

The number .87 converts to 1.1264 in Table 2, and then to 2.1264 by adding 1.0. The number .91 converts to 2.3408 using the same procedure, thus justifying the two numbers obtained in the random sample of size 2. For the stratified sample of size 2, the numbers F(x) = .435 and F(x) = .955 convert to -.1637 + 1 = .8363 and 2.6954 respectively, in agreement with the graphical results but with greater accuracy.

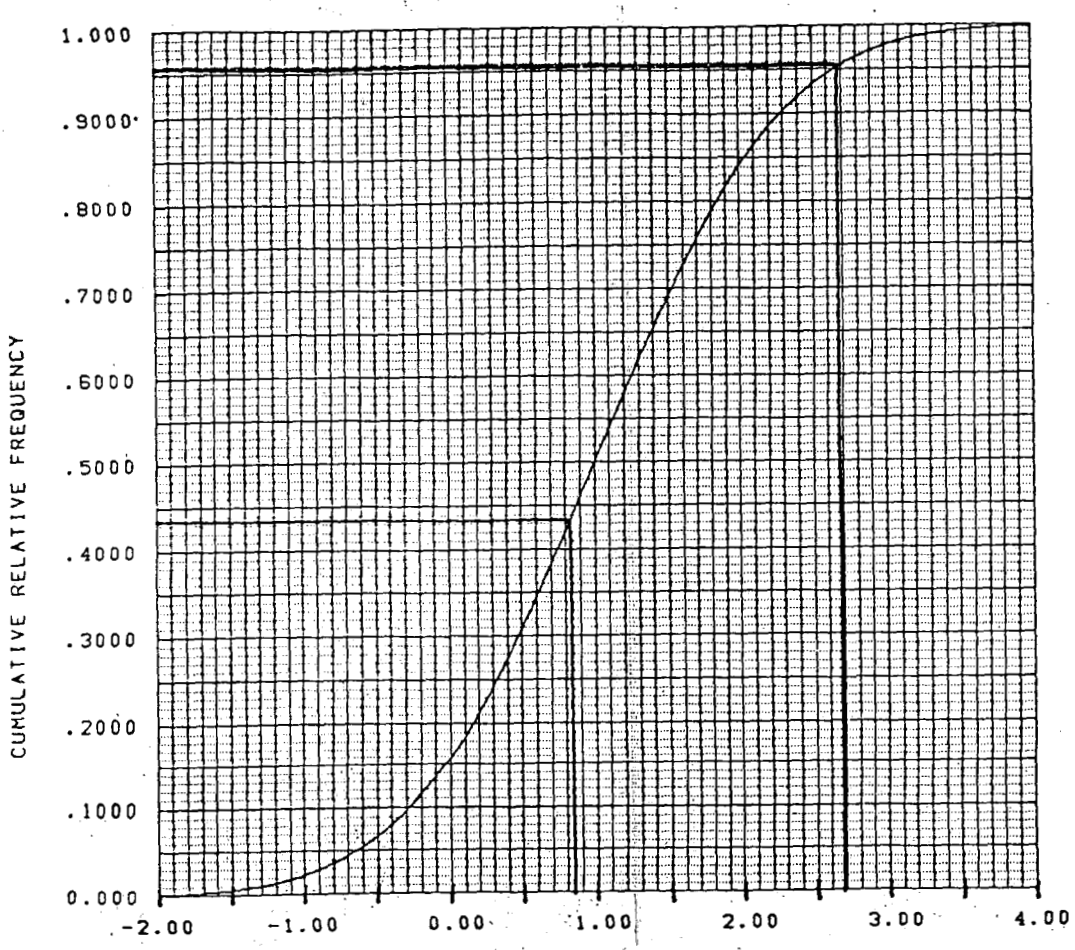


Figure 9. A Hypothetical Distribution Function F(x) for Use in Drawing a Stratified Sample of Size N = 2.

EXERCISE 1

Use Table 1 and Figures 10 and 11 [columns (a) through (d) only], to obtain a random sample of size 4 from the distribution F(x).

(a)	(b) Random numbers (2 digits)	(c) $(c) = \frac{(b)}{100}$	(d) $F^{-1}(c)$	(e) $\Phi^{-1}(c)$	(f) (e) + 1
1.					
2.					
3.					
4.					

Figure 10. Worksheet for Drawing a Random Sample of Size N = 4 from F(x).

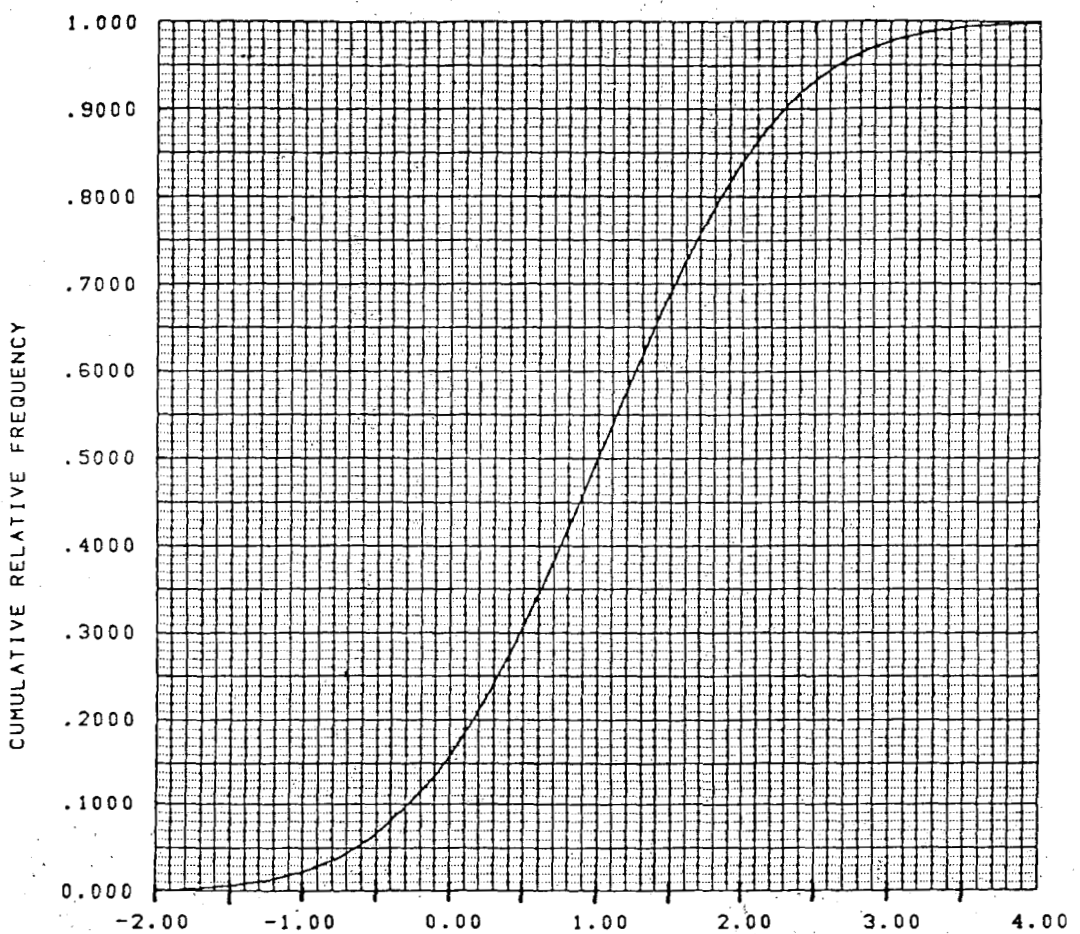


Figure 11. A Hypothetical Distribution Function F(x) for Use in Drawing a Random Sample of Size N = 4.

EXERCISE 2

Use Table 1 and Figures 12 and 13 [columns (a) through (d) only], to obtain a stratified sample of size 4 from the distribution F(x). Note how the stratified sample is spread over all four quarters of the probability distribution.

(a)	(b) Random numbers (2 digits)	(c) $(c) = \frac{(b)}{400} + \frac{(a)-1}{4}$	(d) $F^{-1}(c)$	(e) $\Phi^{-1}(c)$	(f) $\Phi^{-1}(c) + 1$
1.					
2.					
3.					
4.					

Figure 12. Worksheet for Drawing a Latin Hypercube Sample of Size N = 4 from F(x).

QUESTION 1: For purposes of estimating the mean of the population, do you think the average of a random sample or the average of a stratified sample will tend to give a more accurate figure? ANSWER AT BOTTOM OF PAGE 18.

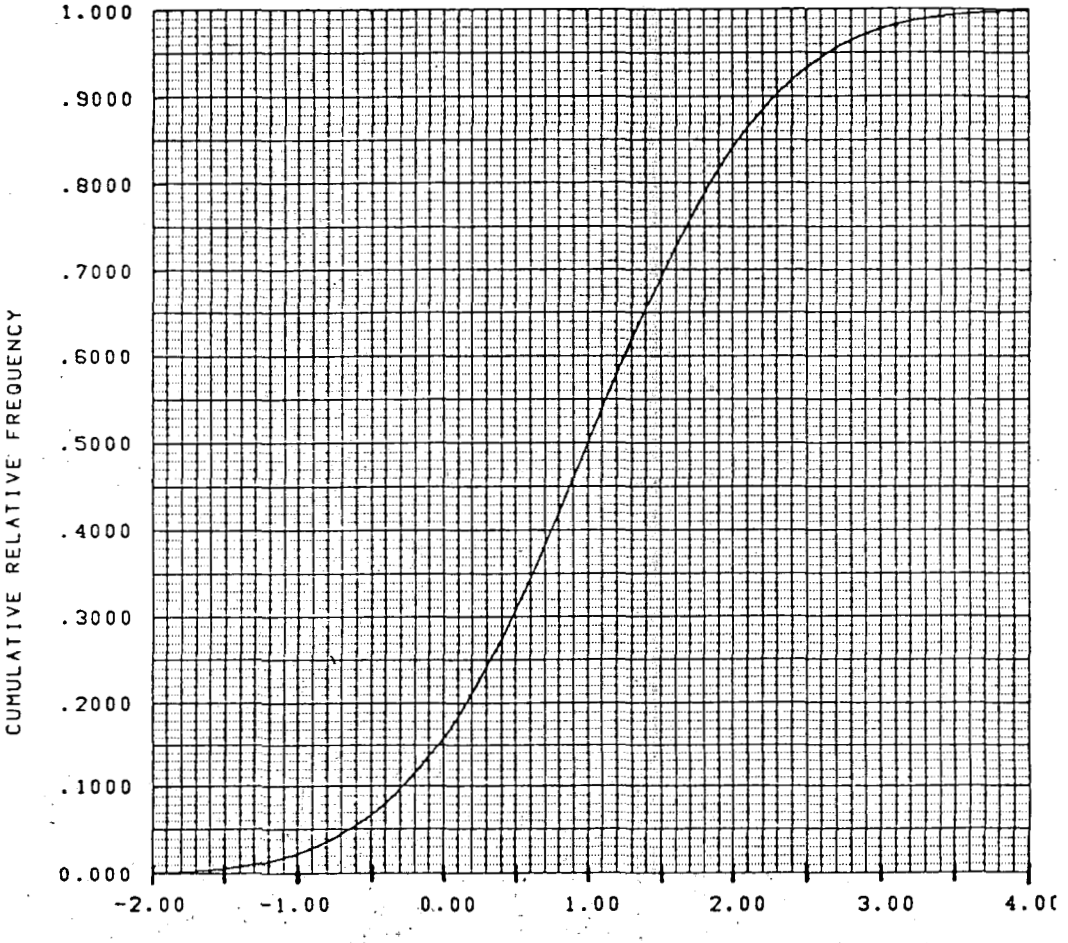


Figure 13. A Hypothetical Distribution Function F(x) for Use in Drawing a Latin Hypercube Sample of Size N = 4.

EXERCISE 3

Use Table 2 and Figure 10, columns (e) and (f), to obtain more accurate figures for the random sample from F(x), the normal distribution with $\mu = 1$ and $\sigma = 1$.

EXERCISE 4

Use Table 2 and Figure 12, columns (e) and (f), to obtain more accurate figures for the stratified sample from F(x).

ANSWER TO QUESTION 1: It can be shown that the average from a stratified sample will tend to be closer to the true population mean.

Obtaining a Multivariate Random Sample

The usual model involves many input variables rather than just the one input variable as was used in the previous examples. If four input variables are involved, as in Figure 2, then one value needs to be obtained for each input variable before an input to the model is complete. Let K be the number of input variables; K = 4 in Figure 2. Then K numbers are obtained as one input vector, where each number represents one input variable. If N is the number of runs desired, then N sets of K numbers each are obtained in all.

The Random Sampling method with K input variables is a simple extension of the Random Sampling method for one input variable, if the input variables are independent. The first observation on X_1 is simply matched with the first observation on X_2, the first observation on X_3, and so on, for the first input vector. The second input vector consists of the second values obtained for X_1, X_2, ..., X_K, and so on for all N input vectors. The situation is not so simple if some specified correlation is desired on the input values, but that will be described later.

Obtaining a Latin Hypercube Sample

To obtain a Latin Hypercube Sample, when the input variables are uncorrelated, the situation is almost as simple. First, stratified samples of size N are obtained on each input variable, in the manner previously described for finding stratified samples. Then the stratified sample of size N on X_1 is permuted into a random order, using some randomization method. The N observations on X_2 are also permuted into a random order, independent of the order on the values of X_1. The values for each input variable are arranged in a random order, independent of the order of the other input variables.

Once the samples are permuted as described, the Latin Hypercube Sample is easily constructed. The first value of X_1 is matched with the first values of X_2, X_3, ..., X_K for the first input vector. The second values of each are matched for the second input vector. This matching procedure is followed until all N values of each variable are used. The method for matching the observations to achieve some target correlation values will be given later, for the non-independent case.

Comparing Latin Hypercube with Random Sampling

A Latin Hypercube Sample has observations that are spread over the entire range of each input variable, and the spread is in a uniform manner, in a probability sense. This is unlike the Random Sample which may produce clusters of observations anywhere in the range of the variables. It is difficult to prove analytically that either method of sampling is better than the other. However, some comparisons of the two sampling methods have been made with actual models, and the Latin

Hypercube Samples appear to give much better results where the goal is to estimate the distribution function of the output variable. The following exercises are designed to illustrate how to obtain Random Samples and Latin Hypercube Samples with multivariate input, when the input variables are uncorrelated. Then comparisons will be made between the two methods to see what kind of accuracy is obtained on estimates of the output.

Obtaining a Multivariate Random Sample (Illustration)

Refer to Figure 2 where there are four input variables each with a normal distribution function. In each run observations on these four input variables go into a model, represented by a black box, and the output is recorded. The following exercise takes the reader step by step through the process of obtaining a random sample of size 10. The subsequent exercise takes the reader through the steps in obtaining a Latin Hypercube Sample of size 10. Later these samples will be entered into a black-box type model and the outputs recorded and compared. But first the samples are obtained.

EXERCISE 5 (Multivariate random sample of size 10)

1. In order to draw a 4-variate random sample of size 10, 40 random numbers are needed from Table 1. Starting where you left off in Exercise 2, choose 40 three-digit random numbers, reading across the table row by row. Record these in the order drawn, down column (b) in Figures 14-17.

2. The numbers in column (b) are converted to probabilities between 0 and 1 by dividing by 1000, for column (c).

3. The probabilities in column (c) are converted to random samples from a standard normal distribution ($\mu = 0$, $\sigma = 1$) by entering Table 2 in the respective row and column, and recording the table entry in column (d).

4. The standard normal values from column (d) are converted to normal values with other means by adding the constant indicated in column (e) of each figure.

5. Transcribe all 40 numbers in column (e) into the respective columns of Figure 18. Each row of Figure 18 represents one input to the black box model.

(a) Obs. Number	(b) Uniform Random Number (3 digits)	(c) $(c) = \frac{(b)}{1000}$	(d) Φ^{-1} (c) from Table 2	(e) Input (e) = (d)+1
1.				
2.				
3.				
4.				
5.				
6.				
7.				
8.				
9.				
10.				

Figure 14. Worksheet for Drawing a Random Sample of Size 10 for X_1 where X_1 is Normal, $\mu = 1$, $\sigma = 1$.

(a) Obs. Number	(b) Uniform Random Number (3 digits)	(c) $(c) = \frac{(b)}{1000}$	(d) Φ^{-1} (c) from Table 2	(e) Input (e) = (d)+2
1.				
2.				
3.				
4.				
5.				
6.				
7.				
8.				
9.				
10.				

Figure 15. Worksheet for Drawing a Random Sample of Size 10 for X_2, where X_2 is Normal, $\mu = 2$, $\sigma = 1$.

(a) Obs. Number	(b) Uniform Random Number	(c) $(c) = \frac{(b)}{1000}$	(d) $\Phi^{-1}(c)$	(e) $(e) = (d) + 2$
1.				
2.				
3.				
4.				
5.				
6.				
7.				
8.				
9.				
10.				

Figure 16. Worksheet for Drawing a Random Sample of Size 10 for X_3, where X_3 is Normal, $\mu = 2$, $\sigma = 1$.

(a) Obs. Number	(b) Uniform Random Number	(c) $(c) = \frac{(b)}{1000}$	(d) $\Phi^{-1}(c)$	(e) $(e) = (d) + 3$
1.				
2.				
3.				
4.				
5.				
6.				
7.				
8.				
9.				
10.				

Figure 17. Worksheet for Drawing a Random Sample of Size 10 for X_4, where X_4 is Normal, $\mu = 3$, $\sigma = 1$.

6. Using a pre-programmed calculator, or the equation

$$Y = X_1 + X_2X_3 - X_2 \ln|X_1| + \exp(X_4/4)$$

whichever is more convenient, find the output value Y for each of the 10 input values (X_1, X_2, X_3, X_4), and record these in column (f) of Figure 18.

7. Plot the 10 values of Y on the abscissa of Figure 19, and draw an empirical distribution function. An empirical distribution function is a step function which equals zero on the left, and proceeding from left to right, rises a height of 1/N at each of the N sample points, until it equals 1.0. This is an estimate of the true distribution function of the output. In this case, start at zero on the left and increase the height of the graph by 1/10 at each observation on Y, as the graph proceeds from left to right. At the largest observed value of Y the graph should jump from a height of .9 to a height of 1.0.

The Accuracy of the Output From a Random Sample

The empirical distribution function in Figure 19 provides an estimate of the population distribution function of the output. The sample mean of the 10 values of Y provides an estimate of the population mean, and other sample values provide estimates of their corresponding population values in the usual manner for random samples.

In order to see how well these samples function as the basis for population estimates, five random samples of size 10 each were obtained using the Latin Hypercube Sampling Program, and were entered into the black box model to obtain outputs. The five empirical distribution functions are given in Figure 20, while Figure 21 presents a picture of the mean of all five e.d.f.'s together. In the background of Figures 20 and 21 is an accurate estimate of the true distribution function of the output, obtained by using a random sample of size N = 1000.

The mean of all five e.d.f.'s, averaged in the vertical direction, is plotted again in Figure 22. This is the same e.d.f. one would obtain if all 50 sample observations were treated as a random sample of size N = 50, which it actually is. Above and below the mean curve in Figure 22 are curves that represent one standard deviation distance, where the standard deviation is computed vertically from the five curves in Figure 20, and smoothed using a three point moving average. The standard deviation is presented to give some idea of the accuracy involved in each individual random sample of size 10.

(a) Obs. Number	(b) Values for X_1	(c) Values for X_2	(d) Values for X_3	(e) Values for X_4	(f) Y (Output)
1.					
2.					
3.					
4.					
5.					
6.					
7.					
8.					
9.					
10.					

Figure 18. The Multivariate Input Vectors Using a Random Sample, and the Corresponding Output.

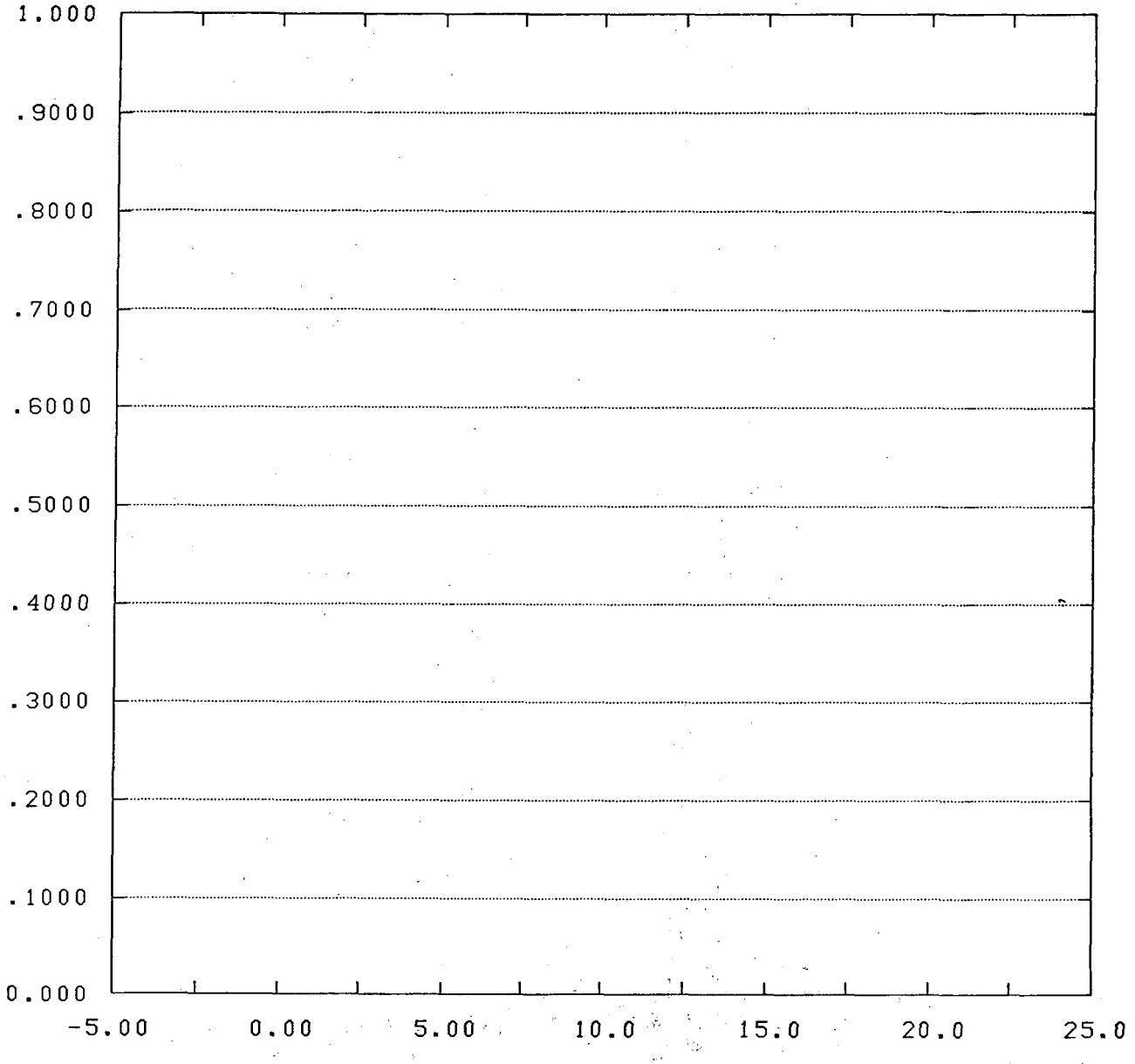


Figure 19. An Empirical Distribution Function from the Random Sample of Figure 18.

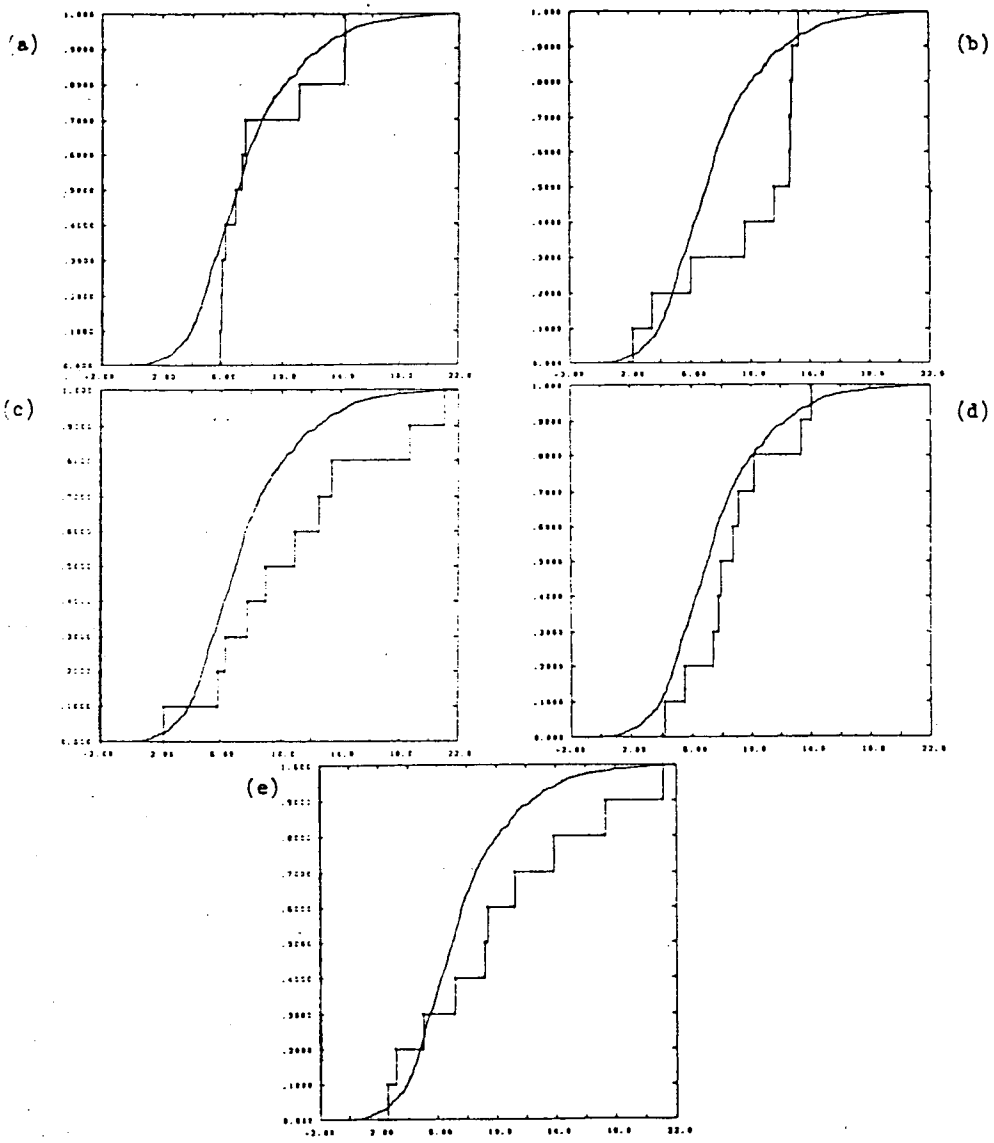


Figure 20. Five Empirical Distribution Function (a) - (e) from Random Samples of Size 10, and an Estimate of the Population Distribution Function

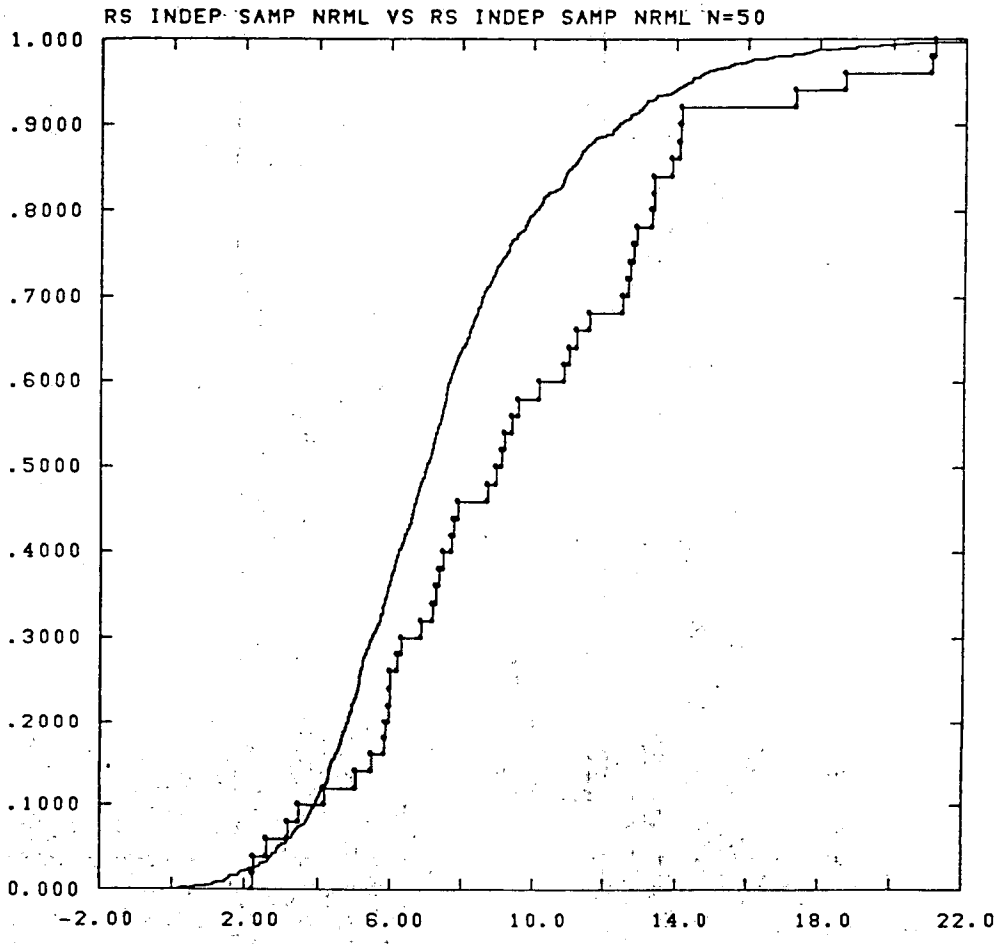


Figure 21. A Graph of the Mean of the Five EDF's from Figure 20, and an Estimate of the Population Distribution Function

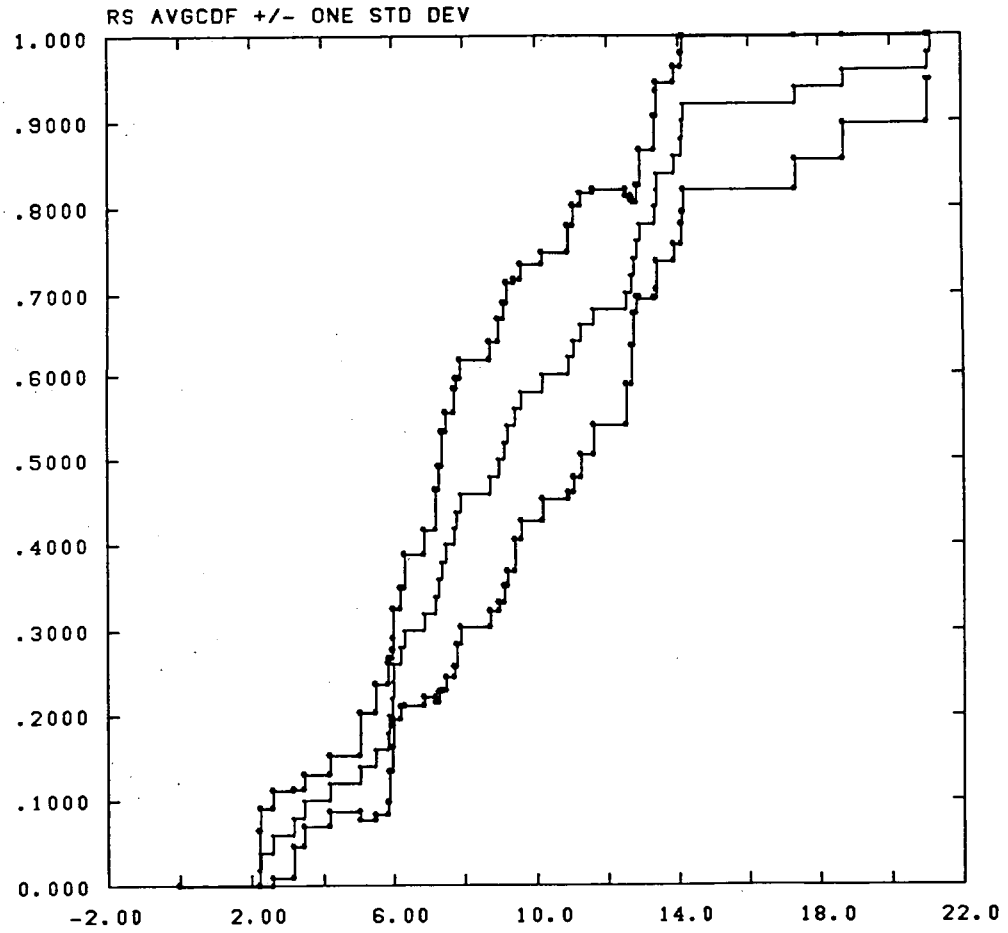


Figure 22. The Mean and One Standard Deviation Bounds of the Five EDFs from Figure 20.

Obtaining a Random Permutation

Before drawing a Latin Hypercube Sample, one method for arranging numbers in a random order will be discussed. One way of obtaining a random permutation of the integers 1 to N is to draw N numbers from a random number table (or a computer program) and use the ranks of those numbers as the random permutation. For example, starting with row 41, column 1 in Table 1, ten consecutive 4-digit numbers are given as follows, along with their ranks.

Random Numbers	Ranks
9842	10
7075	5
2333	2
3626	3
4270	4
0163	1
8924	8
7766	6
9699	9
8420	7

Since the random numbers follow a random ordering, the ranks form a random permutation of the integers from 1 to 10.

Obtaining a Latin Hypercube Sample (Illustration)

The following steps outline the procedure for finding a Latin Hypercube Sample. The situation described in Figure 2 is used. A sample of size N = 10 will be formed using K = 4 input variables, which are independent of each other, and normally distributed with $\sigma^2 = 1$ and means $\mu_1 = 1, \mu_2 = 2, \mu_3 = 2, \mu_4 = 3$ respectively. The reader should follow through the steps, and Figures 23-28 to ensure an understanding of the process.

Step 1. Obtain uniform random numbers in each of N strata. Select a random starting point in Table 1 and, reading across, select 40 two-digit random numbers. Write these in column (b) in Figures 23-26. Divide each of these numbers by 1000 and add the lower bound for the strata from column (a). Put the result in column (c). These are the stratified sample values of F(x).

Step 2. Arrange the values of F(x) in a random order. Draw an additional 40 random numbers from Table 2, using 4-digit numbers to reduce the chances of ties and put these numbers in column (d). In the case of ties, redraw until there are no ties. Rank each 10 of these from 1 to 10 in each Figure and put the ranks in column (e).

(a) Lower Bound of Stratum	(b) Uniform Random Numbers	(c) F(x)	(d) Uniform Random Numbers	(e) Ranks of (d)	(f) Random Order of F(x)	(g) Φ^{-1}	(h) $F^{-1} = (g)+1$
.0	31	.031	7216	9	.854	1.05	2.05
.1	93	.193	3095	2	.193	-0.87	0.13
.2	52	.252	3812	4	.370	-0.33	0.67
.3	70	.370	1510	1	.031	-1.87	-0.87
.4	22	.422	6878	8	.770	0.74	1.74
.5	84	.584	9190	10	.962	1.77	2.77
.6	06	.606	3187	3	.252	-0.67	0.33
.7	70	.770	4934	6	.584	0.21	1.21
.8	54	.854	4055	5	.422	-0.20	0.80
.9	62	.962	6087	7	.606	0.27	1.27

Figure 23. A Stratified Sample of Size 10 for X_1 from a Normal Population with $\mu = 1$ and $\sigma^2 = 1$.

(a) Lower Bound of Stratum	(b) Uniform Random Numbers	(c) F(x)	(d) Uniform Random Numbers	(e) Ranks of (d)	(f) Random Order of F(x)	(g) Φ^{-1}	(h) $F^{-1} = (g)+2$
.0	35	.035	8279	8	.716	0.57	2.57
.1	21	.121	0709	1	.035	-1.81	0.19
.2	61	.261	2565	3	.261	-0.64	1.36
.3	44	.344	3900	4	.344	-0.40	1.60
.4	86	.486	5224	6	.529	0.07	2.07
.5	29	.529	7295	7	.689	0.49	2.49
.6	89	.689	8286	9	.886	1.21	3.21
.7	16	.716	0981	2	.121	-1.17	0.83
.8	86	.886	4063	5	.486	-0.04	1.96
.9	07	.907	9147	10	.907	1.32	3.32

Figure 24. A Stratified Sample of Size 10 for X_2 from a Normal Population with $\mu = 2$ and $\sigma^2 = 1$.

(a) Lower Bound of Stratum	(b) Uniform Random Numbers	(c) F(x)	(d) Uniform Random Numbers	(e) Ranks of (d)	(f) Random Order of F(x)	(g) Φ^{-1}	(h) $F^{-1} = (g)+2$
.0	41	.041	2554	2	.186	-0.89	1.11
.1	86	.186	9485	10	.965	1.81	3.81
.2	71	.271	4242	4	.349	-0.39	1.61
.3	49	.349	6274	7	.669	0.44	2.44
.4	51	.451	5233	5	.451	-0.12	1.88
.5	91	.591	5720	6	.591	0.23	2.23
.6	69	.669	2946	3	.271	-0.61	1.39
.7	68	.768	1723	1	.041	-1.74	0.26
.8	50	.850	7720	8	.768	0.73	2.73
.9	65	.965	7896	9	.850	1.04	3.04

Figure 25. A Stratified Sample of Size 10 for X_3 from a Normal Population with $\mu = 2$ and $\sigma^2 = 1$.

(a) Lower Bound of Stratum	(b) Uniform Random Numbers	(c) F(x)	(d) Uniform Random Numbers	(e) Ranks of (d)	(f) Random Order of F(x)	(g) Φ^{-1}	(h) $F^{-1} = (g)+3$
.0	50	.050	2750	4	.381	-0.30	2.70
.1	00	.100	4961	9	.899	1.28	4.28
.2	13	.213	3183	6	.566	0.17	3.17
.3	81	.381	9444	10	.924	1.43	4.43
.4	40	.440	1575	2	.100	-1.28	1.72
.5	66	.566	1057	1	.050	-1.64	1.36
.6	32	.637	3086	5	.440	-0.15	2.85
.7	11	.711	1964	3	.213	-0.80	2.20
.8	99	.899	4827	8	.711	0.56	3.56
.9	24	.924	3923	7	.637	0.35	3.35

Figure 26. A Stratified Sample of Size 10 for X_4 from a Normal Population with $\mu = 3$ and $\sigma^2 = 1$.

(a) Obs. No.	(b) Input X_1	(c) Input X_2	(d) Input X_3	(e) Input X_4	(f) Output Y
1	2.05	2.57	1.11	2.70	5.02
2	0.13	0.19	3.81	4.28	4.16
3	0.67	1.36	1.61	3.17	5.61
4	-0.87	1.60	2.44	4.43	6.28
5	1.74	2.07	1.88	1.72	6.02
6	2.77	2.49	2.23	1.36	7.19
7	0.33	3.21	1.39	2.85	10.39
8	1.21	0.83	0.26	2.20	3.00
9	0.80	1.96	2.73	3.56	9.02
10	1.27	3.32	3.04	3.35	12.88

Figure 27. A Latin Hypercube Sample of Size 10 and the Associated Output Y.

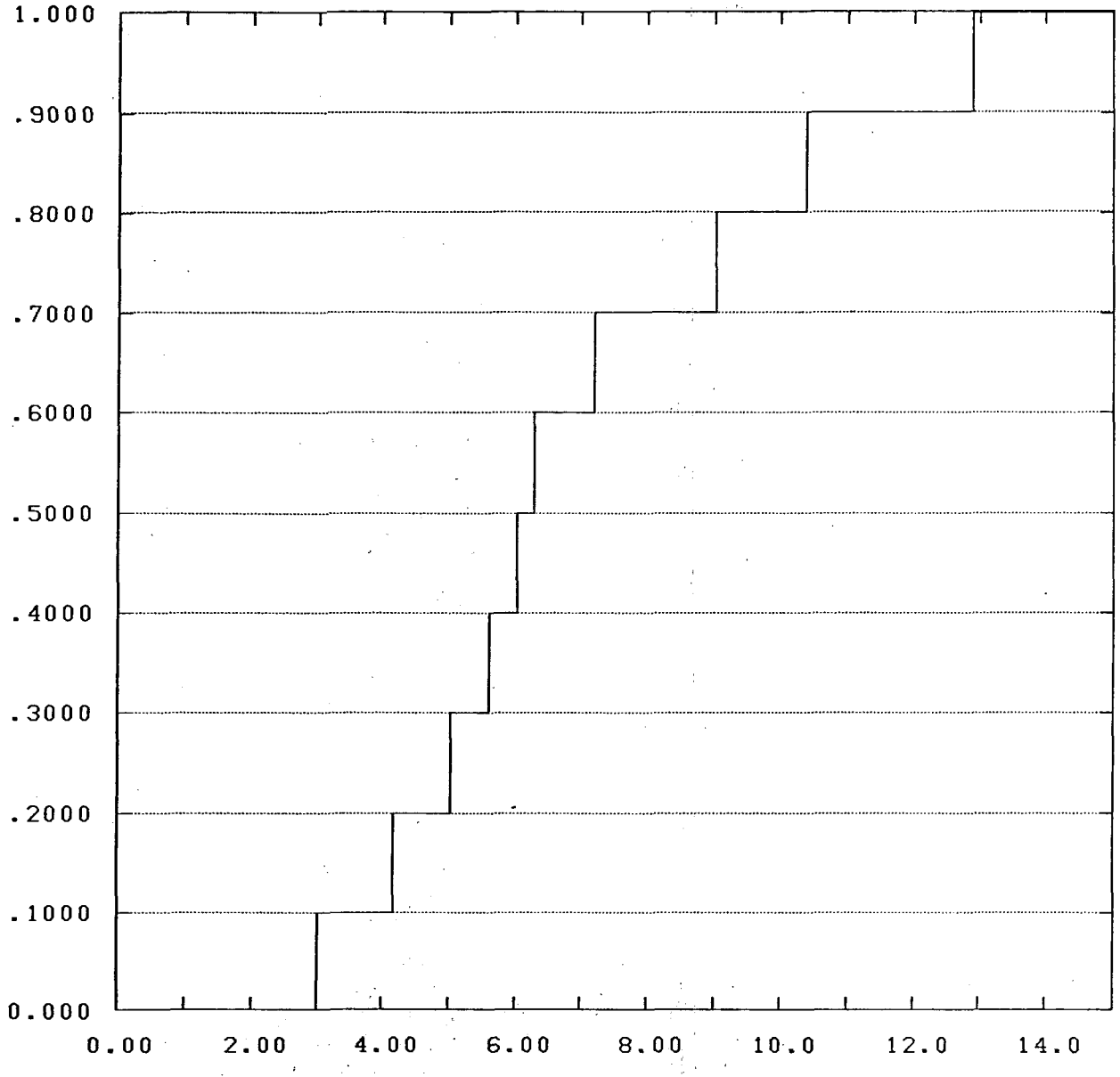


Figure 28. An Empirical Distribution Function for the Output in Figure 27.

Now rearrange the numbers from column (c), putting the smallest number next to rank 1, the second smallest next to rank 2, and so on to the largest number which goes next to rank 10 in each Figure. These go in column (f).

Step 3. Convert the values of F(x) to a stratified sample from F(x). Since F(x) is a normal distribution function, enter the respective row and column of Table 2 as indicated by the value in column (f), and record the table entry Φ^{-1} in column (g). Add the mean of F(x) to column (g) to get the value of F^{-1}, which goes in column (h). Column (h) contains the stratified sample from F(x), arranged in random order.

Step 4. Combine the individual stratified samples into a Latin Hypercube Sample. Transcribe the numbers in column (h) of Figure 23 into column (b) of Figure 27, without changing the relative ordering. In a similar fashion the numbers in column (c) of Figure 27 come from Figure 24, column (d) comes from Figure 25 and column (e) comes from Figure 26. It is important to keep the same relative ordering of the numbers when transcribing them.

Step 5. Obtain the output from the black box model using the Latin Hypercube Sample. The entries in row 1 of Figure 27 are entered into a pre-programmed calculator, or the function

$$Y = X_1 + X_2X_3 - X_2 \ln|X_1| + \exp(X_4/4)$$

whichever is more convenient, to get the output Y of the black box model. Repeat this procedure for each row in Figure 27.

Step 6. Plot an empirical distribution function. Plot the 10 values of Y from Figure 27 onto the horizontal axis of Figure 28. Draw a step function, starting at zero on the left, and increasing in steps of height 1/10 at each value of Y, until the graph reaches a height of 1.0 at the largest value of Y. This is an estimate of the distribution function of the output, obtained using Latin Hypercube Sampling. The average of the 10 values of Y may be used to estimate the population mean; the sample variance, sample median, etc., may be used to estimate the population counterparts.

EXERCISE 6. (Obtaining a Latin Hypercube Sample of size 10)

Follow the same steps used in the previous example, and obtain a Latin Hypercube Sample of size N = 10. Use Figures 29-32 to record the steps involved in finding the stratified samples, and put them together in Figure 33 as a Latin Hypercube Sample. Obtain the output values and graph the empirical distribution function in Figure 34.

(a) Lower Bound of Stratum	(b) Uniform Random Numbers	(c) F(x)	(d) Uniform Random Numbers	(e) Ranks of (d)	(f) Random Order of F(x)	(g) Φ^{-1}	(h) $F^{-1} = (g)+1$
.0							
.1							
.2							
.3							
.4							
.5							
.6							
.7							
.8							
.9							

Figure 29. Student Problem: A Stratified Sample of Size 10 for X_1 from a Normal Population with $\mu = 1$ and $\sigma^2 = 1$.

(a) Lower Bound of Stratum	(b) Uniform Random Numbers	(c) F(x)	(d) Uniform Random Numbers	(e) Ranks of (d)	(f) Random Order of F(x)	(g) Φ^{-1}	(h) $F^{-1} = (g)+2$
.0							
.1							
.2							
.3							
.4							
.5							
.6							
.7							
.8							
.9							

Figure 30. Student Problem: A Stratified Sample of Size 10 for X_2 from a Normal Population with $\mu = 2$ and $\sigma^2 = 1$.

(a) Lower Bound of Stratum	(b) Uniform Random Numbers	(c) F(x)	(d) Uniform Random Numbers	(e) Ranks of (d)	(f) Random Order of F(x)	(g) Φ^{-1}	(h) $F^{-1} = (g)+2$
.0							
.1							
.2							
.3							
.4							
.5							
.6							
.7							
.8							
.9							

Figure 31. Student Problem: A Stratified Sample of Size 10 for X_3 from a Normal Population with $\mu = 2$ and $\sigma^2 = 1$.

(a) Lower Bound of Stratum	(b) Uniform Random Numbers	(c) F(x)	(d) Uniform Random Numbers	(e) Ranks of (d)	(f) Random Order of F(x)	(g) Φ^{-1}	(h) $F^{-1} = (g)+3$
.0							
.1							
.2							
.3							
.4							
.5							
.6							
.7							
.8							
.9							

Figure 32. Student Problem: A Stratified Sample of Size 10 for X_4 from a Normal Population with $\mu = 3$ and $\sigma^2 = 1$.

(a) Obs. No.	(b) Input X_1	(c) Input X_2	(d) Input X_3	(e) Input X_4	(f) Output Y
1					
2					
3					
4					
5					
6					
7					
8					
9					
10					

Figure 33. Student Problem: Worksheet for a Latin Hypercube Sample of Size 10 and the Associated Output Y.

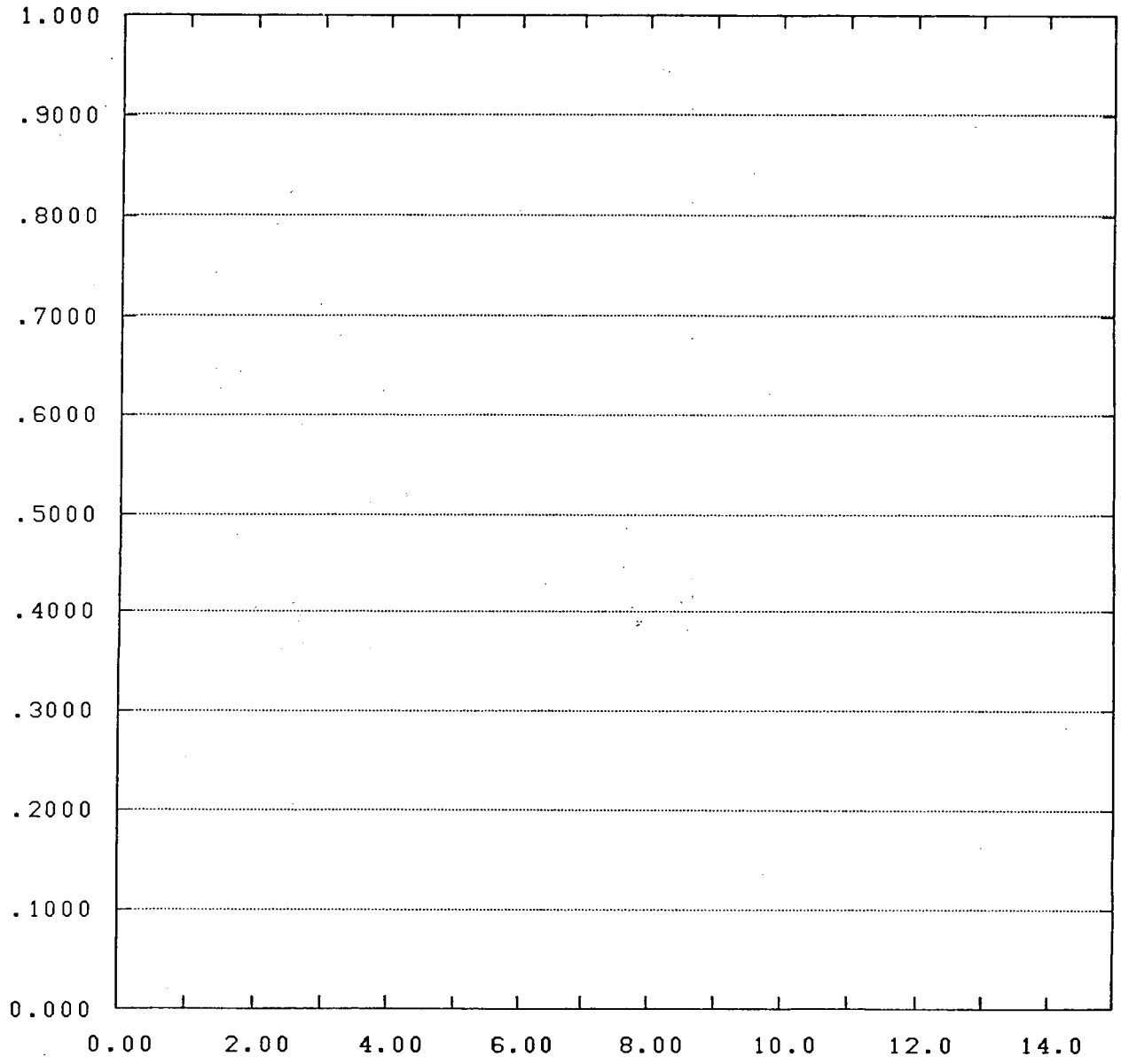


Figure 34. Student Problem: Worksheet for an Empirical Distribution Function for the Output in Figure 33.

Accuracy Obtained from Using a Latin Hypercube Sample

In order to get some idea of how well a Latin Hypercube Sample of size 10 functions as a basis for estimation, the Latin Hypercube Sampling program was used to obtain five Latin Hypercube Samples of size 10 each. The e.d.f.'s for these five samples appear in Figure 35, (a)-(e). The mean of these five graphs, computed in a vertical direction, appears in Figure 36. In the background of Figures 35 and 36 is an estimate of the population distribution function, obtained using a Latin Hypercube Sample of size N = 1000. This estimate coincides almost perfectly with the estimate in Figures 20 and 21 which was obtained from a Random Sample of size N = 1000. This close agreement confirms the fact that both methods of sampling are providing unbiased estimates of the population distribution function.

The mean EDF is plotted again in Figure 37, along with curves plotted one standard deviation above and below the mean curve. These three curves collectively give some idea of the spread involved using Latin Hypercube Samples of size 10 as estimators of the population distribution function. The standard deviation is computed vertically from the five curves in Figure 35, and smoothed using a three point moving average.

A Comparison of Latin Hypercube with Random Sampling

A comparison of Figures 36 and 37 with Figures 21 and 22 shows that, in this case, the five Latin Hypercube Samples provide a better composite estimate of the population distribution function than do the five random samples obtained earlier. Because of sampling variability, there is no guarantee that Latin Hypercube samples are always better than random samples, but all of the simulation studies we are aware of indicate a definite tendency in this direction.

The Replicated Latin Hypercube Sample

When five random samples are pooled together as in Figure 21 the result is another random sample, whose size is equal to the total pooled sample size. However, when five (or any number) Latin Hypercube Samples are pooled together as in Figure 36, the result is called a Replicated Latin Hypercube Sample. Actually the inputs are more correctly called the replicated Latin Hypercube Sample. Replication allows standard deviations to be computed. These standard deviations should be divided by $\sqrt{r}$, where r is the number of replications, to get an estimate of the standard error[1] of the estimate of the mean c.d.f. In the previous example, the standard deviations would be divided by $\sqrt{5}$. The estimate of the true output distribution function obtained from a random sample of size 50, and the standard error of the estimate, is given in Figure 38. This is different than Figure 22 which illustrated the error involved when using a random sample of size 10. The corresponding

[1] Standard error refers to the standard deviation of an estimate

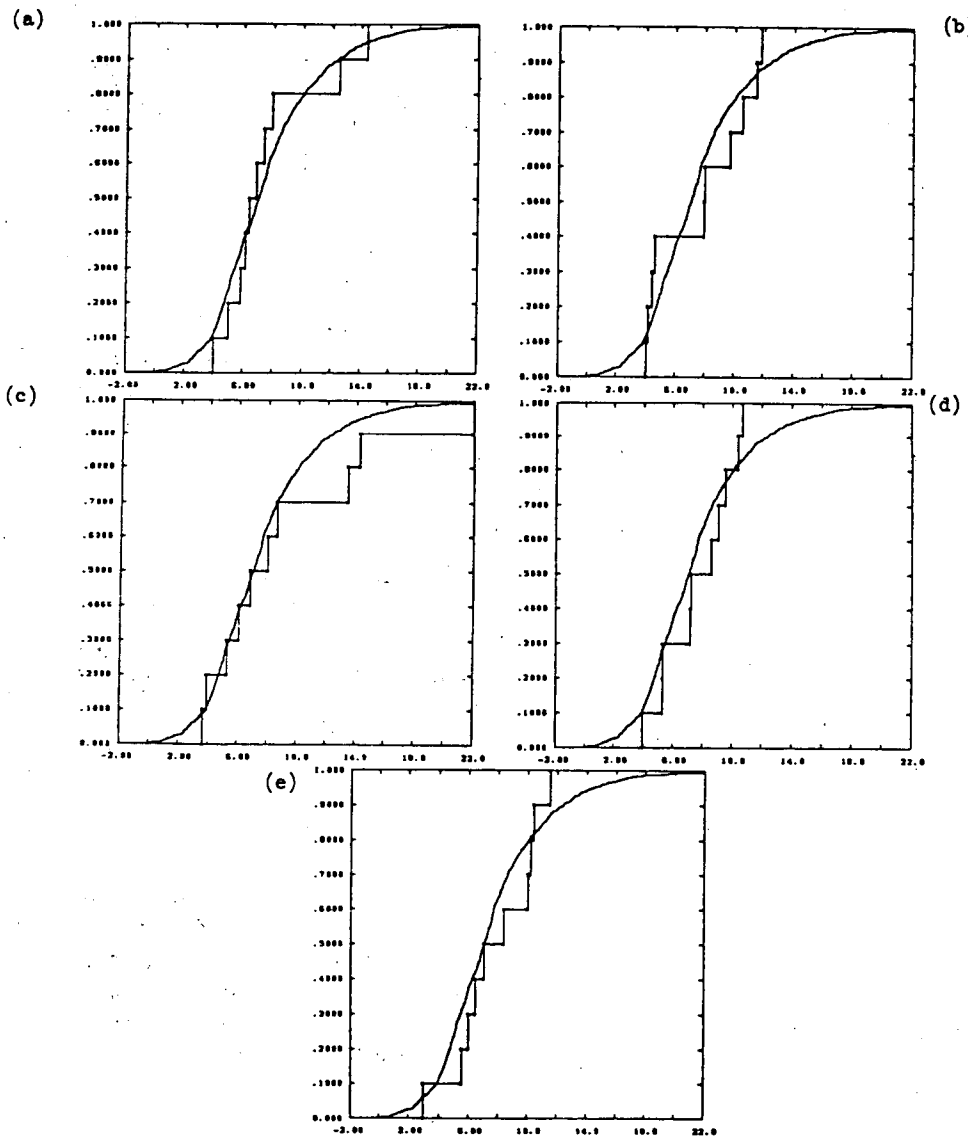


Figure 35. Five EDF's Obtained from Latin Hypercube Samples of Size 10 Each, and an Estimate of the Population Distribution Function

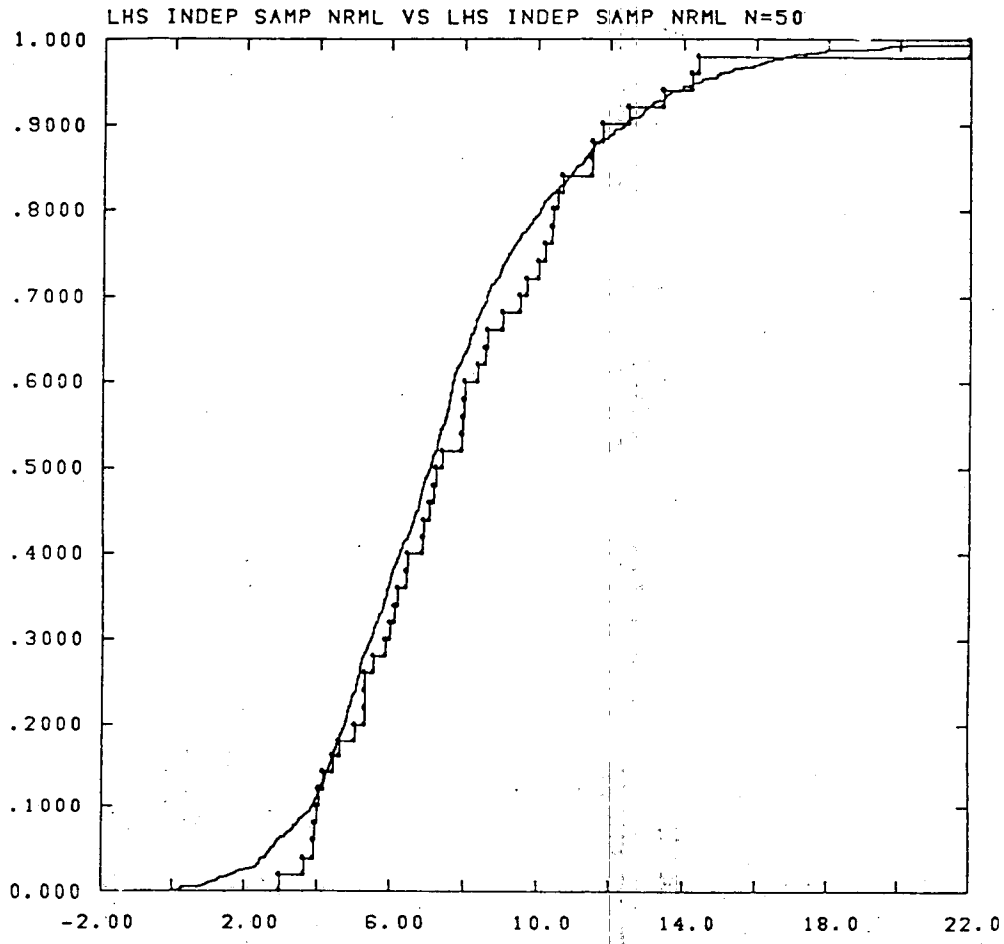


Figure 36. The Mean of the Five EDF's from Figure 35, and an Estimate of the Population Distribution Function

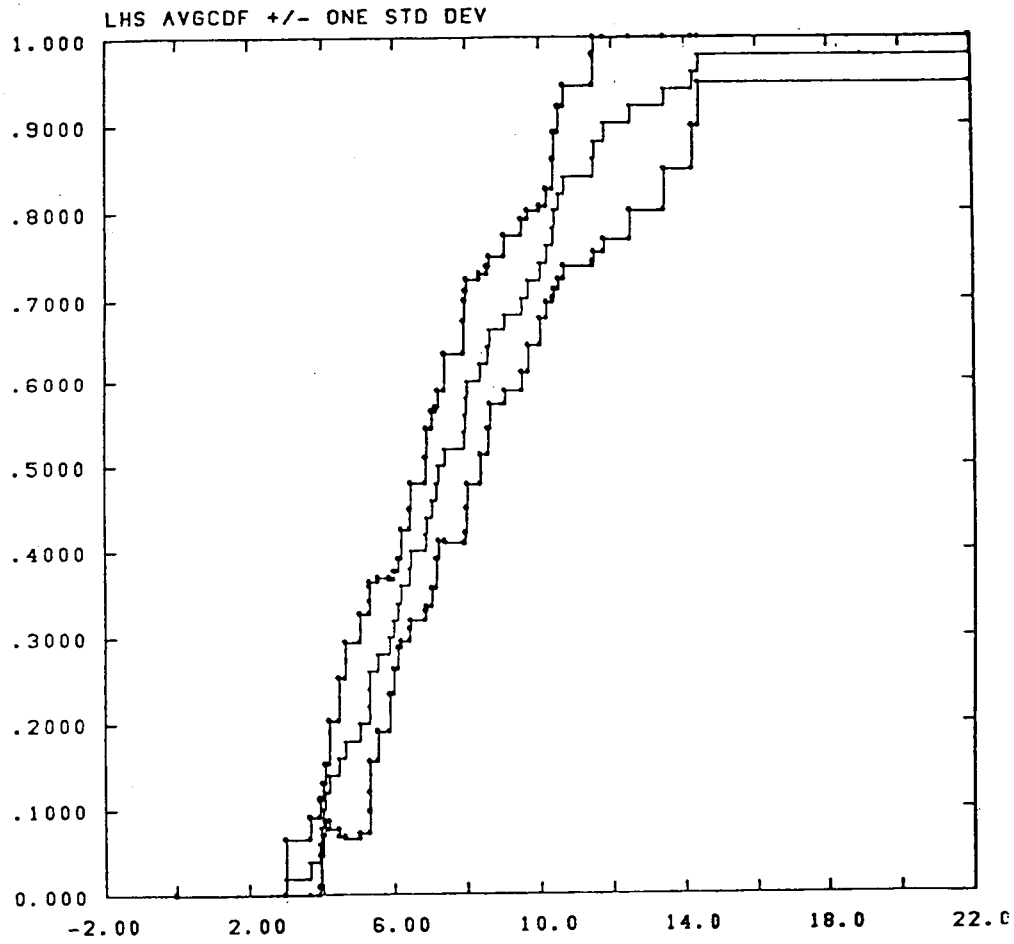


Figure 37. The Mean of the Five EDF's from Figure 35, and One Standard Deviation Bounds

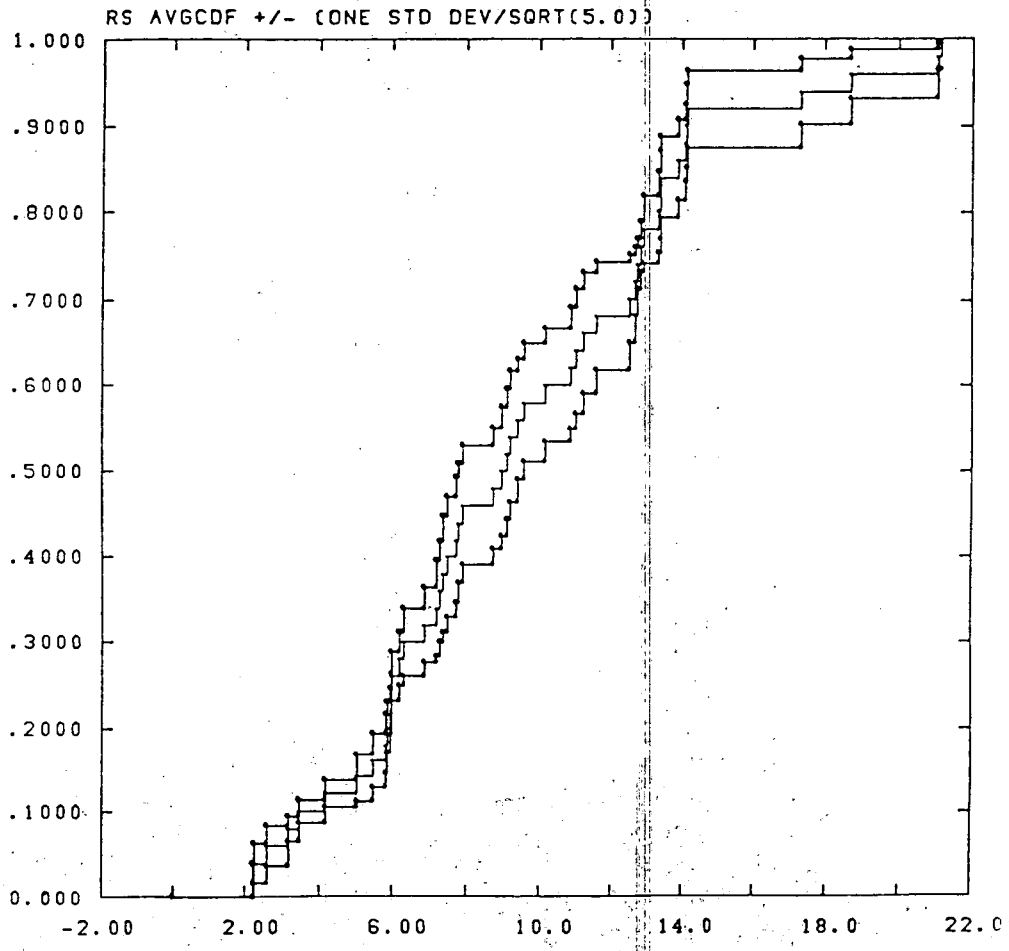


Figure 38. The Mean of Five EDF's from Figure 20 and One Standard Error Bounds for a Random Sample of Size N = 50

figure for a Replicated Latin Hypercube Sample (5 samples of 10 each) is given in Figure 39, showing the estimated distribution function and the one standard error bounds. This may be contrasted with Figure 37 which illustrated the standard error involved when using a single Latin Hypercube Sample of size 10.

Estimating Other Population Parameters

Other population quantities are estimated in the usual way from the sample outputs. For example, the sample mean is used to estimate the population mean. Each of the five random samples provides a sample mean, as does each of the five Latin Hypercube samples. These are listed below.

True population mean $\mu = 7.585$

Random Sample Estimates		Latin Hypercube Sample Estimates	
1.	8.504	1.	7.672
2.	9.736	2.	7.682
3.	10.778	3.	9.266
4.	8.825	4.	7.735
5.	10.029	5.	7.867
ave.	9.575	ave.	8.044

Latin Hypercube Samples appear to provide better estimates of most, if not all, population parameters when compared with Random Samples. However, this observation is based only on empirical evidence, not theoretical proof, and may not be true in particular cases.

Changes in the Input Distributions

If the input distributions are changed, the output distribution will be changed also. Just how much the output distribution function will change depends on the degree of change in the input distributions and the strength of the association between the output and each input variable. For purposes of illustration, the input distributions in the example depicted in Figure 2 were changed from normal to uniform for each of the four variables. The range of each variable remained the same. Now the samples of observations from each distribution will not tend to be in the center of the range, as with a normal distribution, but will tend to be spread evenly from one end of the range to the other. This increased emphasis on values in the tails of the range can be expected to alter the output distribution somewhat, but the degree of change is difficult to predict.

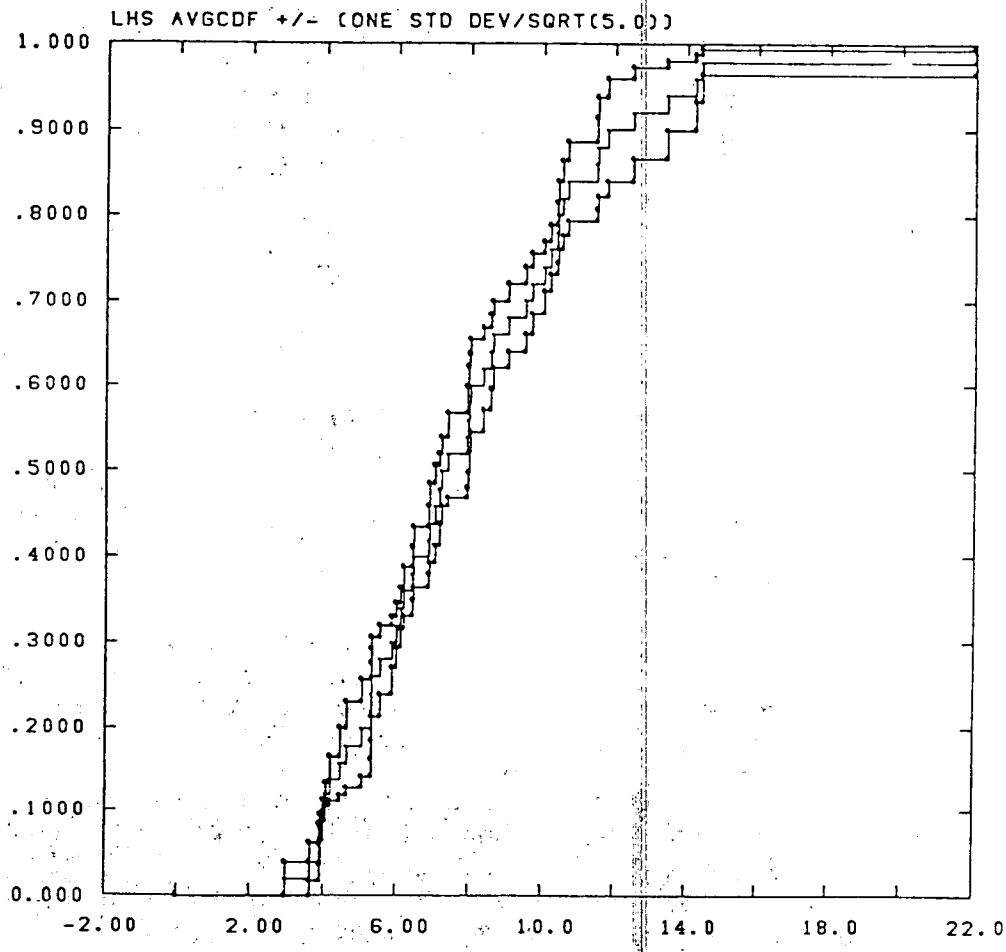


Figure 39. The Mean of Five EDF's from Figure 35 and One Standard Error Bounds for a Replicated Latin Hypercube Sample of Total Size N = 50

Illustrating the Effect of a Change in Input Distributions

To see how much change is induced by changing the four input distributions from normal to uniform, a Latin Hypercube Sample of size N = 1000 was obtained using the Latin Hypercube Sampling Program. The steps involved in finding such a sample are similar to those described earlier. That is, in Figures 23-26, columns (a) through (f) would remain unchanged, but to convert from F(x) to F^{-1}, as given in column (h), the inverse function for a uniform distribution would be used instead of the inverse function for a normal distribution function. The result for a sample of size 1000 is given in Figure 40 (dark line) and contrasted with the previous case involving normal distributions (light line). The large sample size enables these graphs to be treated as if they were the true output distribution functions. The change in the distribution is considerable, which illustrates the importance of being as accurate as possible in specifying the input distribution functions.

The Actual Correlation on the Input Values

Recall that in drawing a multivariate random sample, the process depended on numbers from a random number generator or, in this case, Table 1. Since the numbers drawn in this way are supposed to be independent of one another, any correlation induced should be spurious correlation due simply to usual random fluctuation one might expect to encounter in random samples.

The same is true for the Latin Hypercube Samples, which were dependent on random numbers for the pairing of values of X_1 with X_2 for instance. Since the values of X_1 and X_2 were permuted at random, any correlation between X_1 and X_2 should be spurious correlation. To see how much correlation actually exists between these randomly permuted values, the actual correlation coefficient was computed on the values of X_1 and X_2 given in columns (b) and (c) of Figure 27. That correlation is $r_{12} = .3595$ which is much less than the 5% critical value .632, so a correlation this large can easily be due to chance fluctuations. The entire correlation matrix for the columns in Figure 26 is given in Figure 41.

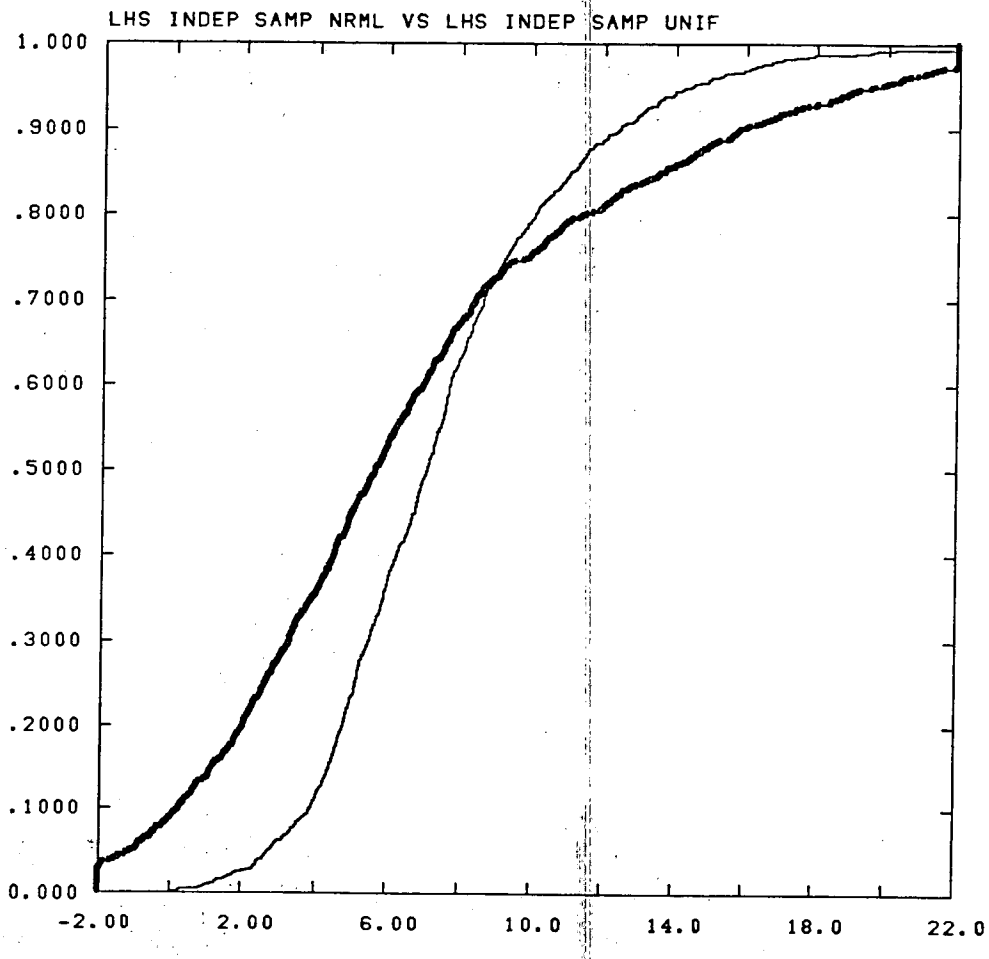


Figure 40. The Output Distribution Function When the Inputs are Uniformly Distributed, Contrasted With the Case of Normal Input Variables.

	X_1	X_2	X_3
X_2	.3595		
X_3	-.2822	-.1024	
X_4	-.8535	-.3170	.5565

Figure 41. The Correlation Matrix for the Latin Hypercube Sample in Figure 27.

Note that five of the six correlation coefficients are nonsignificant at the 5% level. The correlation between X_1 and X_4 is -.8535, which exceeds in absolute value the critical value, but this is merely a chance occurrence, since the sampling method does not induce any systematic correlation in the values. Also, the critical value .632 applies to random samples; the exact critical value for Latin Hypercube Samples is unknown.

The Rank Correlation on the Input Values

Since the exact behavior of the correlation coefficient from Latin Hypercube Samples is not known, and since its behavior even with random sampling is unknown, if the input distributions are not normal, it makes more sense to work with the rank correlation coefficient, known as Spearman's rho, and which is simply r computed on the ranks of the data. The behavior of the rank correlation coefficient is the same for Latin Hypercube Samples as it is for random samples, and is the same for all types of input distributions.

The ranks of the Latin Hypercube Sample of Figure 27 are given in column (e) of Figures 23-26, and are reproduced in Figure 42 for the reader's convenience. The rank correlation matrix for these ranks is given in Figure 43. The 5% critical value for the rank correlation coefficient is .6364. Note that, as before, the only correlation that exceeds this value in absolute value is the rank correlation between X_1 and X_4. Some of the other correlations tend to be large also, such as ρ_{34} = .6121 between X_3 and X_4.

Some Undesirable Effects of Spurious Correlation

These large correlation coefficients that occur by chance after a random permutation of the input variables are annoying for several reasons. For one, the independence assumption between input variables implies that the population correlations equal zero. Since the sample correlations act as estimates of the population values, it would be

Run Number	X_1	X_2	X_3	X_4
1	9	8	2	4
2	2	1	10	9
3	4	3	4	6
4	1	4	7	10
5	8	6	5	2
6	10	7	6	1
7	3	9	3	5
8	6	2	1	3
9	5	5	8	8
10	7	10	9	7

Figure 42. The Ranks of the Input Variables in the Latin Hypercube Sample of Figure 27.

	X_1	X_2	X_3
X_2	.4788		
X_3	-.2848	-.0667	
X_4	-.8061	-.2848	.6121

Figure 43. The Rank Correlation Matrix for the Ranks in Figure 42.

desirable for the sample correlations to be close to zero if possible. In this way, the sample input values would be more "typical" of the population, and this should be reflected in more confidence in the output being more typical of the population output.

A second reason for wanting smaller correlation is that large correlation tends to introduce an effect known as multi-colinearity, which is acceptable if the variables are actually related, but which may be undesireable if the variables are actually independent. For these reasons, a method for reducing sample correlations of input values is desirable. Such a feature is built into the Latin Hpercube Sampling Program.

Reducing the Spurious Correlation

The Latin Hypercube Sampling program does not obtain a random pairing of the input vectors in either the random sampling option or the Latin Hypercube option. Rather, it pairs the variables so they will have correlation coefficients closer to the population correlation coefficients, in order to reduce the undesirable effects associated with spurious correlation. For the first random sample of size 10, whose e.d.f. is given in Figure 20(a), the input variables were arranged so that their ranks matched the ranks given in Figure 44. That is, instead of being satisified with a random ordering such as in column (e) of Figures 14-17, the values are arranged so that their ordering agrees with the ordering in Figure 44. Then the rank correlations are given in Figure 45. Note that the largest rank correlation in Figure 45 is .2217, and that four of the six correlations are less than .1. These correlations as a group tend to be closer to the zero population value that is appropriate for independent input variables. Note also that the rank correlations in Figure 45 depend only on the ranks in Figure 44, and not in any way on the input distributions or the particular input values.

An Illustration of Reducing the Correlation

The same rank ordering given in Figure 44 was used on both the first random sample and the first Latin Hypercube Sample, whose e.d.f. is given in Figure 35(a). To illustrate how this is accomplished, the example given in Figures 23-27 will be reworked so that the correlation matrix of the input values will be the same as in Figure 45. This means that the ranks of the input values need to agree with the columns of Figure 44. For variable X_1 the original ordering and the new ordering are given in Figure 46. The original ordering was given in Figure 23, column (h) and the original ranks were given in the same figure, column (e). The new rank ordering for X_1 is given in Figure 44, column (b). Since the new ordering has rank 8 in run number 1, the X_1 value with rank 8, $X_1 = 1.74$, is now listed first. All of the values of X_1 are thus arranged to agree with the new rankings, as illustrated in Figure 46.

(a) Run Number	(b) X_1	(c) X_2	(d) X_3	(e) X_4
1	8	6	1	9
2	7	7	10	5
3	2	2	6	6
4	4	5	3	7
5	9	10	5	3
6	6	4	9	1
7	1	9	2	2
8	10	3	4	8
9	5	1	7	4
10	3	8	8	10

Figure 44. The Rank Ordering Induced on the Random Sample of Size N = 10, Whose Output EDF is Given in Figure 20(a).

	X_1	X_2	X_3
X_2	-.0944		
X_3	.0938	.0200	
X_4	.2217	-.0252	-.2046

Figure 45. The Rank Correlation Coefficients for the Ranks in Figure 44.

(a) Run Number	(b) Original Ordering Fig. 23, col. (h)	(c) Original Ranks Fig. 23 col. (e)	(d) New Ranks Fig. 44 col. (b)	(e) New Ordering
1	2.05	9	8	1.74
2	0.13	2	7	1.27
3	0.67	4	2	0.13
4	-0.87	1	4	0.67
5	1.74	8	9	2.05
6	2.77	10	6	1.21
7	0.33	3	1	-0.87
8	1.21	6	10	2.77
9	0.80	5	5	0.80
10	1.27	7	3	0.33

Figure 46. Changing the Order of the Values of X_1 to Reduce the Spurious Correlation.

Similarly the values of X_2 are arranged in the ordering suggested by the ranks in column (c) of Figure 44. Columns (d) and (e) of Figure 44 define the new orderings of the values of X_3 and X_4. When all of the input variables are arranged in the order specified by Figure 44, their rank correlation matrix will be the same as the one in Figure 45. Of course, the output values Y will not be the same as before since the combinations of input values are now different. However, these new output values are treated the same as any other output values; e.d.f.'s are plotted as they were in Figures 20 and 35, sample means and sample standard deviations are computed, and so on.

Simulating Correlated Input Variables

The same principle that is used to make the sample correlations close to zero is used by the Latin Hypercube Sampling program to make the sample correlation close to any target correlation. That is, first the pairings of ranks are found that result in a desired sample rank correlation coefficient, and then the sample values are arranged in the order suggested by the ranks. These sample correlation coefficients will not equal exactly their target correlations, just as the sample correlations in the previous example did not equal exactly zero, but they will usually be fairly close. The Latin Hypercube Sampling program furnishes a matrix of ranks to use for the ordering of the sample values, and also furnishes the sample correlation matrix associated with those ranks. If the sample correlation matrix is unsatisfactory for any reason, the program can be used again and again to furnish new rank orderings until a rank ordering with a satisfactory sample rank correlation matrix is found. (See Iman and Conover, 1980).

Illustration of Correlating Input Variables

Suppose the target correlation matrix is given by Figure 47. These values are supplied to the Latin Hypercube Sampling program, and a matrix of ranks that may be used is supplied. One such matrix is given in Figure 48. Note that any sample of size 10 for each of four input values may be arranged in the order suggested by Figure 48. It does not matter if a random sample is used, or if a Latin Hypercube Sample is used, or what types of input distributions are used. When the sample values have the ordering of Figure 48, they will have the sample rank correlation coefficients given in Figure 49.

The sample values from the Latin Hypercube sample given in Figure 27 are rearranged in the order suggested by Figure 48. First the X_1 values are rearranged so that their rank ordering is changed from the former order, 9, 2, 4, 1, 8, 10, 3, 6, 5, 7, as given in column (e) of Figure 23, to the new order 1, 9, 8, 3, 7, 4, 2, 6, 5, 10, as given in column (b) of Figure 48. This is the same type of procedure that was illustrated in Figure 46, only the new ordering is different, since now the objective is to achieve correlations close to zero, as was formerly the case.

	X_1	X_2	X_3
X_2	.8		
X_3	.3	.4	
X_4	.6	.9	.7

Figure 47. A Target Correlation Matrix for Four Input Variables.

(a) Run Number	(b) X_1	(c) X_2	(d) X_3	(e) X_4
1	1	2	2	3
2	9	10	9	10
3	8	8	8	8
4	3	3	10	6
5	7	9	7	9
6	4	7	3	7
7	2	1	6	2
8	6	5	4	5
9	5	4	1	1
10	10	6	5	4

Figure 48. Rank Orderings to Achieve Correlations Close to Those in Figure 47.

	X_1	X_2	X_3
X_2	.7939		
X_3	.3212	.3576	
X_4	.5152	.8545	.6606

Figure 49. Sample Rank Correlation Matrix for Ranks in Figure 48.

After rearranging the order of the sample values for X_1 through X_4 into the order suggested by the ranks given in Figure 48, the new arrangements are given in Figure 50. The values given in columns (b) through (e) of Figure 50 are obtained from column (h) of Figures 23 through 26 respectively. The ranks in Figure 50 are the ranks in Figure 48, so the sample rank correlation matrix of the data given in Figure 50 is given by Figure 49.

Note that even though this example happened to use a Latin Hypercube Sample, a random sample could have been used just as well. All that is required is that the values have the same ordering as given in Figure 48, and the sample rank correlation matrix will be the one in Figure 49.

A New Output Distribution Function

For the sake of illustration, the new combinations of input values shown in Figure 50 were entered into the black box model. The output values are listed in column (f) of Figure 50. These are different output values than those given in Figure 27, because the input values occur in different combinations than before. The e.d.f. is therefore different than before. This is as it should be because the population distribution function being estimated is different than it was before. That is, the true output distribution function depends on what correlations the input variables have, as well as what the input distributions are. The difference in the output distributions due to the correlations structure, Figure 47, being assumed rather than assuming independence of the input variables is shown by the difference in the two curves in Figure 51. The darker curve in Figure 51 is the true output distribution when the inputs are correlated, and the lighter curve is the one which results from independent inputs. Because of this difference, it is important for the input variables to simulate the population correlation matrix, through the use of some device such as the one built into the Latin Hypercube Sampling program.

How Many Runs are Needed?

One of the main advantages of the Latin hypercube sampling procedure is that the number of runs can be very small, regardless of the number of variables involved. In fact, there is no lower limit to the number of runs (input vectors) if the user is willing to sacrifice some of the statistical analyses that are available with larger numbers of runs.

To be able to use the correlation reduction techniques, the number of runs needs to exceed the number of variables. This procedure is more stable if the number of runs exceeds the number of variables by approximately 25%. This is also a minimum requirement for stability in the regression procedures and partial correlation coefficients described in the next part of this tutorial. Of course, more runs than this results in even more stability, and for best results, the number of runs should be about two or three times the number of variables involved if time and money permit.

(a) Run Number	(b) Input X_1	(rank)	(c) Input X_2	(rank)	(d) Input X_3	(rank)	(e) Input X_4	(rank)	(f) Output Y
1	-0.87	(1)	0.83	(2)	1.11	(2)	2.20	(3)	1.90
2	2.05	(9)	3.32	(10)	3.04	(9)	4.43	(10)	12.79
3	1.74	(8)	2.57	(8)	2.73	(8)	3.56	(8)	9.77
4	0.33	(3)	1.36	(3)	3.81	(10)	3.17	(6)	9.23
5	1.27	(7)	3.21	(9)	2.44	(7)	4.28	(9)	11.25
6	0.67	(4)	2.49	(7)	1.39	(3)	3.35	(7)	7.44
7	0.13	(2)	0.19	(1)	2.23	(6)	1.72	(2)	2.48
8	1.21	(6)	1.96	(5)	1.61	(4)	2.85	(5)	6.03
9	0.80	(5)	1.60	(4)	0.26	(1)	1.36	(1)	2.98
10	2.77	(10)	2.07	(6)	1.88	(5)	2.70	(4)	6.52

Figure 50. New Input Values with Rank Correlation Matrix Given by Figure 49.

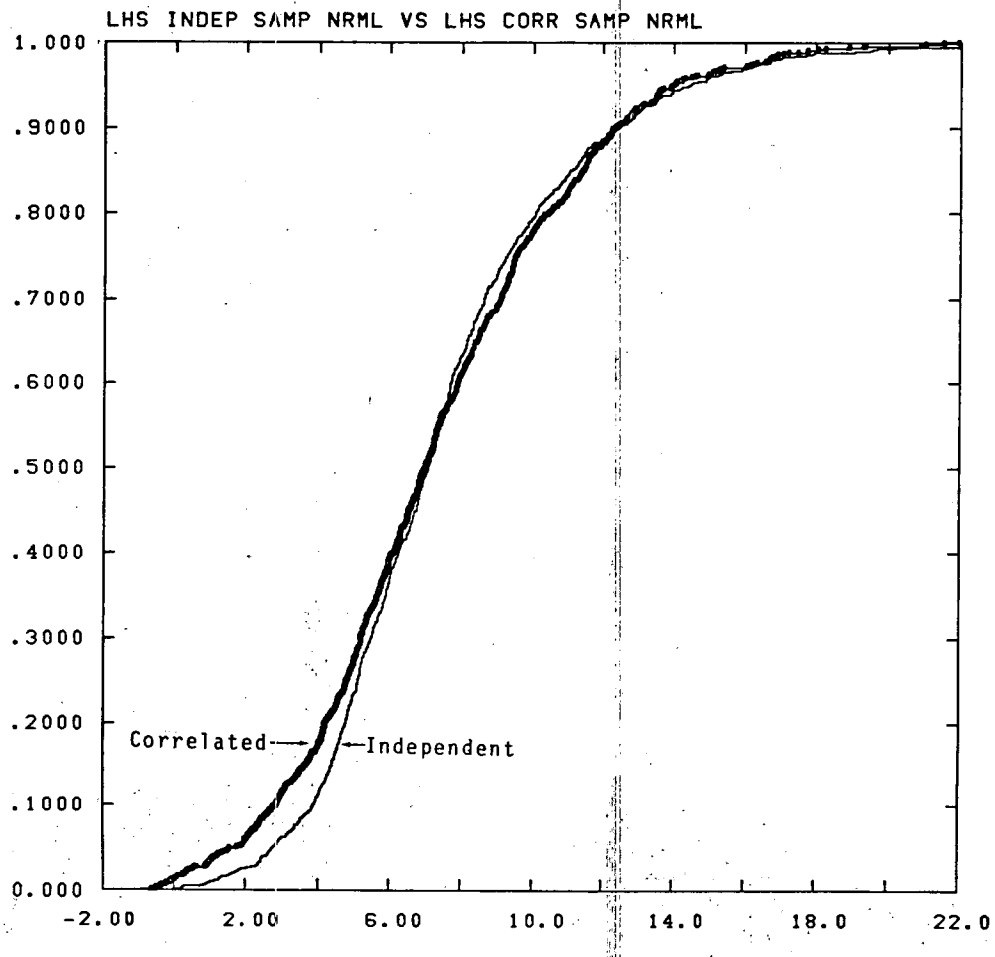


Figure 51. An Example of the Differences in Output Distributions Obtained, Assuming Input Variable Independence and Assuming a Correlation Between Input Variables.

A TABLE OF 14,000 RANDOM UNITS

Line/Col.	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)	(12)	(13)	(14)
1	10480	15011	01536	02011	81647	91646	69179	14194	62590	36207	20969	99570	91291	90700
2	22368	46573	25595	85393	30995	89198	27982	53402	93965	34095	52666	19174	39615	99505
3	24130	48360	22527	97265	76393	64809	15179	24830	49340	32081	30680	19655	63348	58629
4	42167	93093	06243	61680	07856	16376	39440	53537	71341	57004	00849	74917	97758	16379
5	37570	39975	81837	16656	06121	91782	60468	81305	49684	60672	14110	06927	01263	54613
6	77921	06907	11008	42751	27756	53498	18602	70659	90655	15053	21916	81825	44394	42880
7	99562	72905	56420	69994	98872	31016	71194	18738	44013	48840	63213	21069	10634	12952
8	96301	91977	05463	07972	18876	20922	94595	56869	69014	60045	18425	84903	42508	32307
9	89579	14342	63661	10281	17453	18103	57740	84378	25331	12566	58678	44947	05585	56941
10	85475	36857	43342	53988	53060	59533	38867	62300	08158	17983	16439	11458	18593	64952
11	28918	69578	88231	33276	70997	79936	56865	05859	90106	31595	01547	85590	91610	78188
12	63553	40961	48235	03427	49626	69445	18663	72695	52180	20847	12234	90511	33703	90322
13	09429	93969	52636	92737	88974	33488	36320	17617	30015	08272	84115	27156	30613	74952
14	10365	61129	87529	85689	48237	52267	67689	93394	01511	26358	85104	20285	29975	89868
15	07119	97336	71048	08178	77233	13916	47564	81056	97735	85977	29372	74461	28551	90707
16	51085	12765	51821	51259	77452	16308	60756	92144	49442	53900	70960	63990	75601	40719
17	02368	21382	52404	60268	89368	19885	55322	44819	01188	65255	64835	44919	05944	55157
18	01011	54092	33362	94904	31273	04146	18594	29852	71585	85030	51132	01915	92747	64951
19	52162	53916	46369	58586	23216	14513	83149	98736	23495	64350	94738	17752	35156	35749
20	07056	97628	33787	09998	42698	06691	76988	13602	51851	46104	88916	19509	25625	58104
21	48663	91245	85828	14346	09172	30168	90229	04734	59193	22178	30421	61666	99904	32812
22	54164	58492	22421	74103	47070	25306	76468	26384	58151	06646	21524	15227	96909	44592
23	32639	32363	05597	24200	13363	38005	94342	28728	35806	06912	17012	64161	18296	22851
24	29334	27001	87637	87308	58731	00256	45834	15398	46557	41135	10367	07684	36188	18510
25	02488	33062	28834	07351	19731	92420	60952	61280	50001	67658	32586	86679	50720	94953
26	81525	72295	04839	96423	24878	82651	66566	14778	76797	14780	13300	87074	79666	95725
27	29676	20591	68086	26432	46901	20849	89768	81536	86645	12659	92259	57102	80428	25280
28	00742	57392	39064	66432	84673	40027	32832	61362	98947	96067	64760	64584	96096	98253
29	05366	04213	25669	26422	44407	44048	37937	63904	45766	66134	75470	66520	34693	90449
30	91921	26418	64117	94305	26766	25940	39972	22209	71500	64568	91402	42416	07844	69618
31	00582	04711	87917	77341	42206	35126	74087	99547	81817	42607	43808	76655	62028	76630
32	00725	69884	62797	56170	86324	88072	76222	36086	84637	93161	76038	65855	77919	88006
33	69011	65797	95876	55293	18988	27354	26575	08625	40801	59920	29841	80150	12777	48501
34	25976	57948	29888	88604	67917	48708	18912	82271	65424	69774	33611	54262	85963	03547
35	09763	83473	73577	12908	30883	18317	28290	35797	05998	41698	34952	37888	38917	88050
36	91567	42595	27958	30134	04024	86385	29880	99730	55536	84855	29080	09250	79656	73211
37	17955	56349	90999	49127	20044	59931	06115	20542	18059	02008	73708	83517	36103	42791
38	46503	18584	18845	49618	02304	51038	20655	58727	28168	15475	56942	53389	20562	87338
39	92157	89634	94824	78171	84610	82834	09922	25417	44137	48413	25555	21246	35509	20468
40	14577	62765	35605	81263	39667	47358	56873	56307	61607	49518	89656	20103	77490	18062
41	98427	07523	33362	64270	01638	92477	66969	98420	04880	45585	46565	04102	46880	45709
42	34914	63976	88720	82765	34476	17032	87589	40836	32427	70002	70663	88863	77775	69348
43	70060	28277	39475	46473	23219	53416	94970	25832	69975	94884	19661	72828	00102	66794
44	53976	54914	06990	67245	68350	82948	11398	42878	80287	88267	47363	46634	06541	97809
45	76072	29515	40980	07391	58745	25774	22987	80059	39911	96189	41151	14222	60697	59583
46	90725	52210	83974	29992	65831	38857	50490	83765	55657	14361	31720	57375	56228	41546
47	64364	67412	33339	31926	14883	24413	59744	92351	97473	89286	35931	04110	23726	51900
48	08962	00358	31662	25388	61642	34072	81249	35648	56891	69352	48373	45578	78547	81788
49	95012	68379	93526	70765	10593	04542	76463	54328	02349	17247	28865	14777	62730	92277
50	15664	10493	20492	38391	91132	21999	59516	81652	27195	48223	46751	22923	32261	85653

A TABLE OF 14,000 RANDOM UNITS

Line/Col.	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)	(12)	(13)	(14)
51	16408	81899	04153	53381	79401	21438	83035	92350	36693	31238	59649	91754	72772	02338
52	18629	81953	05520	91962	04739	13092	97662	24822	94730	06496	35090	04822	86772	98289
53	73115	35101	47498	87637	99016	71060	88824	71013	18735	20286	23153	72924	35165	43040
54	57491	16703	23167	49323	45021	33132	12544	41035	80780	45393	44812	12515	98931	91202
55	30405	83946	23792	14422	15059	45799	22716	19792	09983	74353	68668	30429	70735	25499
56	16631	35006	85900	98275	32388	52390	16815	69298	82732	38480	73817	32523	41961	44437
57	96773	20206	42559	78985	05300	22164	24369	54224	35083	19687	11052	91491	60383	19746
58	38935	64202	14349	82674	66523	44133	00697	35552	35970	19124	63318	29686	03387	59846
59	31624	76384	17403	53363	44167	64486	64758	75366	76554	31601	12614	33072	60332	92325
60	78919	19474	23632	27889	47914	02584	37680	20801	72152	39339	34806	08930	85001	87820
61	03931	33309	57047	74211	63445	17361	62825	39908	05607	91284	68833	25570	38818	46920
62	74426	33278	43972	10119	89917	15665	52872	73823	73144	88662	88970	74492	51805	99378
63	09066	00903	20795	95452	92648	45454	09552	88815	16553	51125	79375	97596	16296	66092
64	42238	12426	87025	14267	20979	04508	64535	31355	86064	29472	47689	05974	52468	16834
65	16153	08002	26504	41744	81959	65642	74240	56302	00033	67107	77510	70625	28725	34191
66	21457	40742	29820	96783	29400	21840	15035	34537	33310	06116	95240	15957	16572	06004
67	21581	57802	02050	89728	17937	37621	47075	42080	97403	48626	68995	43805	33386	21597
68	55612	78095	83197	33732	05810	24813	86902	60397	16489	03264	88525	42786	05269	92532
69	44657	66999	99324	51281	84463	60563	79312	93454	68876	25471	93911	25650	12682	73572
70	91340	84979	46949	81973	37949	61023	43997	15263	80644	43942	89203	71795	99533	50501
71	91227	21199	31935	27022	84067	05462	35216	14486	29891	68607	41867	14951	91696	85065
72	50001	38140	66321	19924	72163	09538	12151	06878	91903	18749	34405	56087	82790	70925
73	65390	05224	72958	28609	81406	39147	25549	48542	42627	45233	57202	94617	23772	07896
74	27504	96131	83944	41575	10573	08619	64482	73923	36152	05184	94142	25299	84387	34925
75	37169	94851	39117	89632	00959	16487	65536	49071	39782	17095	02330	74301	00275	48280
76	11508	70225	51111	38351	19444	66499	71945	05422	13442	78675	84081	66938	93654	59894
77	37449	30362	06694	54690	04052	53115	62757	95348	78662	11163	81651	50245	34971	52924
78	46515	70331	85922	38329	57015	15765	97161	17869	45349	61796	66345	81073	49106	79860
79	30986	81223	42416	58353	21532	30502	32305	86482	05174	07901	54339	58861	74818	46942
80	63798	64995	46583	09765	44160	78128	83991	42865	92520	83531	80377	35909	81250	54238
81	82486	84846	99254	67632	43218	50076	21361	64816	51202	88124	41870	52689	51275	83556
82	21885	32906	92431	09060	64297	51674	64126	62570	26123	05155	59194	52799	28225	85762
83	60336	98782	07408	53458	13564	59089	26445	29789	85205	41001	12535	12133	14645	23541
84	43937	46891	24010	25560	86355	33941	25786	54990	71899	15475	95434	98227	21824	19585
85	97656	63175	89303	16275	07100	92063	21942	18611	47348	20203	18534	03862	78095	50136
86	03299	01221	05418	38982	55758	92237	26759	86367	21216	98442	08303	56613	91511	75928
87	79626	06486	03574	17668	07785	76020	79924	25651	83325	88428	85076	72811	22717	50585
88	85636	68335	47539	03129	65651	11977	02510	26113	99447	68645	34327	15152	55230	93448
89	18039	14367	61337	06177	12143	46609	32989	74014	64708	00533	35398	58408	13261	47908
90	08362	15656	60627	36478	65648	16764	53412	09013	07832	41574	17639	82163	60859	75567
91	79556	29068	04142	16268	15387	12856	66227	38358	22478	73373	88732	09443	82558	05250
92	92608	82674	27072	32534	17075	27698	98204	63863	11051	34648	88022	56148	34925	57031
93	23982	25835	40055	67006	12293	02753	14827	22235	35071	99704	37543	11601	35503	85171
94	09915	96306	05908	97901	28395	14186	00821	80703	70426	75647	76310	88717	37890	40129
95	50937	33300	26695	62247	69927	76123	50842	43834	86654	70959	79725	93872	28117	19233
96	42488	78077	69882	61657	34136	79180	97526	43092	04098	73571	80799	76536	71255	64239
97	46764	86273	63003	93017	31204	36692	40202	35275	57306	55543	53203	18098	47625	88684
98	03237	45430	55417	63282	90816	17349	88298	90183	36600	78406	06216	95787	42579	90730
99	86591	81482	52667	61583	14972	90053	89534	76036	49199	43716	97548	04379	46370	28672
100	38534	01715	94964	87288	65680	43772	39560	12918	86537	62738	19636	51132	25739	56947

	.000	.001	.002	.003	.004	.005	.006	.007	.008	.009
.00		-3.0902	-2.8782	-2.7478	-2.6521	-2.5758	-2.5121	-2.4573	-2.4089	-2.3656
.01	-2.3263	-2.2904	-2.2571	-2.2262	-2.1973	-2.1701	-2.1444	-2.1201	-2.0969	-2.0749
.02	-2.0537	-2.0335	-2.0141	-1.9954	-1.9774	-1.9600	-1.9431	-1.9268	-1.9110	-1.8957
.03	-1.8808	-1.8663	-1.8522	-1.8384	-1.8250	-1.8119	-1.7991	-1.7866	-1.7744	-1.7624
.04	-1.7507	-1.7392	-1.7279	-1.7169	-1.7060	-1.6954	-1.6849	-1.6747	-1.6646	-1.6546
.05	-1.6449	-1.6352	-1.6258	-1.6164	-1.6072	-1.5982	-1.5893	-1.5805	-1.5718	-1.5632
.06	-1.5548	-1.5464	-1.5382	-1.5301	-1.5220	-1.5141	-1.5063	-1.4985	-1.4909	-1.4833
.07	-1.4758	-1.4684	-1.4611	-1.4538	-1.4466	-1.4395	-1.4325	-1.4255	-1.4187	-1.4118
.08	-1.4051	-1.3984	-1.3917	-1.3852	-1.3787	-1.3722	-1.3658	-1.3595	-1.3532	-1.3469
.09	-1.3408	-1.3346	-1.3285	-1.3225	-1.3165	-1.3106	-1.3047	-1.2988	-1.2930	-1.2873
.10	-1.2816	-1.2759	-1.2702	-1.2646	-1.2591	-1.2536	-1.2481	-1.2426	-1.2372	-1.2319
.11	-1.2265	-1.2212	-1.2160	-1.2107	-1.2055	-1.2004	-1.1952	-1.1901	-1.1850	-1.1800
.12	-1.1750	-1.1700	-1.1650	-1.1601	-1.1552	-1.1503	-1.1455	-1.1407	-1.1359	-1.1311
.13	-1.1264	-1.1217	-1.1170	-1.1123	-1.1077	-1.1031	-1.0985	-1.0939	-1.0893	-1.0848
.14	-1.0803	-1.0758	-1.0714	-1.0669	-1.0625	-1.0581	-1.0537	-1.0494	-1.0450	-1.0407
.15	-1.0364	-1.0322	-1.0279	-1.0237	-1.0194	-1.0152	-1.0110	-1.0069	-1.0027	-.9986
.16	-.9945	-.9904	-.9863	-.9822	-.9782	-.9741	-.9701	-.9661	-.9621	-.9581
.17	-.9542	-.9502	-.9463	-.9424	-.9385	-.9346	-.9307	-.9269	-.9230	-.9192
.18	-.9154	-.9116	-.9078	-.9040	-.9002	-.8965	-.8927	-.8890	-.8853	-.8816
.19	-.8779	-.8742	-.8705	-.8669	-.8633	-.8596	-.8560	-.8524	-.8488	-.8452
.20	-.8416	-.8381	-.8345	-.8310	-.8274	-.8239	-.8204	-.8169	-.8134	-.8099
.21	-.8064	-.8030	-.7995	-.7961	-.7926	-.7892	-.7858	-.7824	-.7790	-.7756
.22	-.7722	-.7688	-.7655	-.7621	-.7588	-.7554	-.7521	-.7488	-.7454	-.7421
.23	-.7388	-.7356	-.7323	-.7290	-.7257	-.7225	-.7192	-.7160	-.7128	-.7095
.24	-.7063	-.7031	-.6999	-.6967	-.6935	-.6903	-.6871	-.6840	-.6808	-.6776
.25	-.6745	-.6713	-.6682	-.6651	-.6620	-.6588	-.6557	-.6526	-.6495	-.6464
.26	-.6433	-.6403	-.6372	-.6341	-.6311	-.6280	-.6250	-.6219	-.6189	-.6158
.27	-.6128	-.6098	-.6068	-.6038	-.6008	-.5978	-.5948	-.5918	-.5888	-.5858
.28	-.5828	-.5799	-.5769	-.5740	-.5710	-.5681	-.5651	-.5622	-.5592	-.5563
.29	-.5534	-.5505	-.5476	-.5446	-.5417	-.5388	-.5359	-.5330	-.5302	-.5273
.30	-.5244	-.5215	-.5187	-.5158	-.5129	-.5101	-.5072	-.5044	-.5015	-.4987
.31	-.4959	-.4930	-.4902	-.4874	-.4845	-.4817	-.4789	-.4761	-.4733	-.4705
.32	-.4677	-.4649	-.4621	-.4593	-.4565	-.4538	-.4510	-.4482	-.4454	-.4427
.33	-.4399	-.4372	-.4344	-.4316	-.4289	-.4261	-.4234	-.4207	-.4179	-.4152
.34	-.4125	-.4097	-.4070	-.4043	-.4016	-.3989	-.3961	-.3934	-.3907	-.3880
.35	-.3853	-.3826	-.3799	-.3772	-.3745	-.3719	-.3692	-.3665	-.3638	-.3611
.36	-.3585	-.3558	-.3531	-.3505	-.3478	-.3451	-.3425	-.3398	-.3372	-.3345
.37	-.3319	-.3292	-.3266	-.3239	-.3213	-.3186	-.3160	-.3134	-.3107	-.3081
.38	-.3055	-.3029	-.3002	-.2976	-.2950	-.2924	-.2898	-.2871	-.2845	-.2819
.39	-.2793	-.2767	-.2741	-.2715	-.2689	-.2663	-.2637	-.2611	-.2585	-.2559
.40	-.2533	-.2508	-.2482	-.2456	-.2430	-.2404	-.2378	-.2353	-.2327	-.2301
.41	-.2275	-.2250	-.2224	-.2198	-.2173	-.2147	-.2121	-.2096	-.2070	-.2045
.42	-.2019	-.1993	-.1968	-.1942	-.1917	-.1891	-.1866	-.1840	-.1815	-.1789
.43	-.1764	-.1738	-.1713	-.1687	-.1662	-.1637	-.1611	-.1586	-.1560	-.1535
.44	-.1510	-.1484	-.1459	-.1434	-.1408	-.1383	-.1358	-.1332	-.1307	-.1282
.45	-.1257	-.1231	-.1206	-.1181	-.1156	-.1130	-.1105	-.1080	-.1055	-.1030
.46	-.1004	-.0979	-.0954	-.0929	-.0904	-.0878	-.0853	-.0828	-.0803	-.0778
.47	-.0753	-.0728	-.0702	-.0677	-.0652	-.0627	-.0602	-.0577	-.0552	-.0527
.48	-.0502	-.0476	-.0451	-.0426	-.0401	-.0376	-.0351	-.0326	-.0301	-.0276
.49	-.0251	-.0226	-.0201	-.0175	-.0150	-.0125	-.0100	-.0075	-.0050	-.0025

Table 2. The Cumulative Standard Normal Distribution

	.000	.001	.002	.003	.004	.005	.006	.007	.008	.009
.50	.0000	.0025	.0050	.0075	.0100	.0125	.0150	.0175	.0201	.0226
.51	.0251	.0276	.0301	.0326	.0351	.0376	.0401	.0426	.0451	.0476
.52	.0502	.0527	.0552	.0577	.0602	.0627	.0652	.0677	.0702	.0728
.53	.0753	.0778	.0803	.0828	.0853	.0878	.0904	.0929	.0954	.0979
.54	.1004	.1030	.1055	.1080	.1105	.1130	.1156	.1181	.1206	.1231
.55	.1257	.1282	.1307	.1332	.1358	.1383	.1408	.1434	.1459	.1484
.56	.1510	.1535	.1560	.1586	.1611	.1637	.1662	.1687	.1713	.1738
.57	.1764	.1789	.1815	.1840	.1866	.1891	.1917	.1942	.1968	.1993
.58	.2019	.2045	.2070	.2096	.2121	.2147	.2173	.2198	.2224	.2250
.59	.2275	.2301	.2327	.2353	.2378	.2404	.2430	.2456	.2482	.2508
.60	.2533	.2559	.2585	.2611	.2637	.2663	.2689	.2715	.2741	.2767
.61	.2793	.2819	.2845	.2871	.2898	.2924	.2950	.2976	.3002	.3029
.62	.3055	.3081	.3107	.3134	.3160	.3186	.3213	.3239	.3266	.3292
.63	.3319	.3345	.3372	.3398	.3425	.3451	.3478	.3505	.3531	.3558
.64	.3585	.3611	.3638	.3665	.3692	.3719	.3745	.3772	.3799	.3826
.65	.3853	.3880	.3907	.3934	.3961	.3989	.4016	.4043	.4070	.4097
.66	.4125	.4152	.4179	.4207	.4234	.4261	.4289	.4316	.4344	.4372
.67	.4399	.4427	.4454	.4482	.4510	.4538	.4565	.4593	.4621	.4649
.68	.4677	.4705	.4733	.4761	.4789	.4817	.4845	.4874	.4902	.4930
.69	.4959	.4987	.5015	.5044	.5072	.5101	.5129	.5158	.5187	.5215
.70	.5244	.5273	.5302	.5330	.5359	.5388	.5417	.5446	.5476	.5505
.71	.5534	.5563	.5592	.5622	.5651	.5681	.5710	.5740	.5769	.5799
.72	.5828	.5858	.5888	.5918	.5948	.5978	.6008	.6038	.6068	.6098
.73	.6128	.6158	.6189	.6219	.6250	.6280	.6311	.6341	.6372	.6403
.74	.6433	.6464	.6495	.6526	.6557	.6588	.6620	.6651	.6682	.6713
.75	.6745	.6776	.6808	.6840	.6871	.6903	.6935	.6967	.6999	.7031
.76	.7063	.7095	.7128	.7160	.7192	.7225	.7257	.7290	.7323	.7356
.77	.7388	.7421	.7454	.7488	.7521	.7554	.7588	.7621	.7655	.7688
.78	.7722	.7756	.7790	.7824	.7858	.7892	.7926	.7961	.7995	.8030
.79	.8064	.8099	.8134	.8169	.8204	.8239	.8274	.8310	.8345	.8381
.80	.8416	.8452	.8488	.8524	.8560	.8596	.8633	.8669	.8705	.8742
.81	.8779	.8816	.8853	.8890	.8927	.8965	.9002	.9040	.9078	.9116
.82	.9154	.9192	.9230	.9269	.9307	.9346	.9385	.9424	.9463	.9502
.83	.9542	.9581	.9621	.9661	.9701	.9741	.9782	.9822	.9863	.9904
.84	.9945	.9986	1.0027	1.0069	1.0110	1.0152	1.0194	1.0237	1.0279	1.0322
.85	1.0364	1.0407	1.0450	1.0494	1.0537	1.0581	1.0625	1.0669	1.0714	1.0758
.86	1.0803	1.0848	1.0893	1.0939	1.0985	1.1031	1.1077	1.1123	1.1170	1.1217
.87	1.1264	1.1311	1.1359	1.1407	1.1455	1.1503	1.1552	1.1601	1.1650	1.1700
.88	1.1750	1.1800	1.1850	1.1901	1.1952	1.2004	1.2055	1.2107	1.2160	1.2212
.89	1.2265	1.2319	1.2372	1.2426	1.2481	1.2536	1.2591	1.2646	1.2702	1.2759
.90	1.2816	1.2873	1.2930	1.2988	1.3047	1.3106	1.3165	1.3225	1.3285	1.3346
.91	1.3408	1.3469	1.3532	1.3595	1.3658	1.3722	1.3787	1.3852	1.3917	1.3984
.92	1.4051	1.4118	1.4187	1.4255	1.4325	1.4395	1.4466	1.4538	1.4611	1.4684
.93	1.4758	1.4833	1.4909	1.4985	1.5063	1.5141	1.5220	1.5301	1.5382	1.5464
.94	1.5548	1.5632	1.5718	1.5805	1.5893	1.5982	1.6072	1.6164	1.6258	1.6352
.95	1.6449	1.6546	1.6646	1.6747	1.6849	1.6954	1.7060	1.7169	1.7279	1.7392
.96	1.7507	1.7624	1.7744	1.7866	1.7991	1.8119	1.8250	1.8384	1.8522	1.8663
.97	1.8808	1.8957	1.9110	1.9268	1.9431	1.9600	1.9774	1.9954	2.0141	2.0335
.98	2.0537	2.0749	2.0969	2.1201	2.1444	2.1701	2.1973	2.2262	2.2571	2.2904
.99	2.3263	2.3656	2.4089	2.4573	2.5121	2.5758	2.6521	2.7478	2.8782	3.0902

Table 2 (continued) The Cumulative Standard Normal Distribution

TUTORIAL ON THE REGRESSION PROGRAM

The Purpose of the Course

This is a tutorial on regression methods. It introduces and discusses the topics of regression on one variable, rank regression on one variable, and then proceeds to the case of regression on several variables, using raw data or ranks. Stepwise regression is discussed as a means of selecting important variables. Other regression procedures known as forward regression and backward regression are also mentioned. At the conclusion of this tutorial the reader should be able to understand better the regression program described in "Stepwise Regression with PRESS and Rank Regression (Program Users Guide)" by Iman, Davenport, Frost and Shortencarier (1980).

The Need for Regression Methods

Regression methods are useful for identifying and/or defining the relationship between a variable of interest Y and one or more observable variables, called independent variables and denoted by X_1, X_2, etc. Although regression methods are used in a variety of ways, their primary importance on the study of geologic models is for identifying the input variables X_1, X_2, ... which are the most influential on the output variable Y. For this specific goal of identification of important variables, some regression methods are particularly useful. These include rank regression and stepwise regression. To lead into these topics, simple regression is introduced first.

Simple Linear Regression

Simple regression refers to the case where only one independent variable is considered. Observations (y_1, x_1), (y_2, x_2), ..., on the bivariate random variable (Y, X) are analyzed to see what type of relationship may exist between Y and X. In all but artificial situations, the exact relationship between Y and X cannot be expressed mathematically, so regression methods are directed toward approximating the exact relationship between Y and X with mathematical equations. The simpliest mathematical equation is a straight line, so a straight line is the most popular equation to fit to a set of points such as the observations (y_1, x_1), (y_2, x_2), ..., (y_n, x_n). The equation of a straight line is

$$y = a + bx$$

The analysis using linear regression begins with finding estimates $\hat{a}$ and $\hat{b}$ so that the straight line will agree well with the data. The most popular method of finding $\hat{a}$ and $\hat{b}$ is called the method of least squares.

The Method of Least Squares

When an estimated regression equation is obtained, the observed values of X may be substituted into the regression equation to get values of Y which are called predicted values of Y and are denoted by $\hat{Y}$. In the case

of simple linear regression the predicted values are given by

$$\hat{Y} = \hat{a} + \hat{b}x$$

where $\hat{a}$ and $\hat{b}$ are sample estimates. The least squares method for finding $\hat{a}$ and $\hat{b}$ chooses the numbers that minimize the sum of squares

$$SS = \Sigma(Y - \hat{Y})^2$$

This method assures that the predicted values $\hat{Y}$ will be as close as possible (in the least squares sense) to the observed values Y.

The Least Squares Equations

In this case of one independent variable the least squares solutions for $\hat{a}$ and $\hat{b}$ are simple to express

$$\hat{b} = \frac{\sum_{i=1}^{n} x_i y_i - (\Sigma x_i)(\Sigma y_i)/n}{\sum_{i=1}^{n} x_i^2 - (\Sigma x_i)^2/n} \qquad (1)$$

$$\hat{a} = \frac{1}{n}(\Sigma y_i - b\Sigma x_i) \qquad (2)$$

and may be computed on a hand calculator. When there are two or more independent variables the calculations become much more difficult and are usually performed on a computer.

Example

A simple example is used to illustrate the computations involved in the least squares solution to simple linear regression. The same example will then be used to introduce rank regression. Suppose five observations on (Y, X) are obtained as given in Figure 1. The least squares coefficients are found using Equations (1) and (2).

$$\hat{b} = \frac{199.83 - (29.0)(28.4)/5}{189.94 - (28.4)^2/5} = \frac{35.110}{28.628} = 1.226$$

$$\hat{a} = \frac{1}{5}(29.0 - (1.226)(28.4)) = -1.164$$

A graph of the five observations and the least squares regression line

$$\hat{Y} = -1.164 + 1.226X$$

are given in Figure 2.

Obs. Pair	Y	X	X^2	XY
1	1.4	2.3	5.29	3.22
2	5.3	4.1	16.81	21.73
3	4.8	5.6	31.36	26.88
4	6.5	7.2	51.84	46.80
5	11.0	9.2	84.64	101.20
Total	29.0	28.4	189.94	199.83

Figure 1. Worksheet for Finding $\hat{a}$ and $\hat{b}$ Using Least Squares.

The linear regression model appears to fit the points in Figure 2 fairly well. The residuals $(Y - \hat{Y})$ are a measure of how well the linear regression model agrees with the data points

$$\text{residual} = Y - \hat{Y} = Y - (-1.164 + 1.226X)$$

This choice of coefficients, $\hat{a} = -1.164$ and $\hat{b} = 1.226$, results in the smallest possible sum of squares achievable using a straight line to fit the data. In this case the minimum value is SS = 5.0803, as given in Figure 3.

Obs. Pair	Y	X	$\hat{Y}$	Residual	$(Y-\hat{Y})^2$
1	1.4	2.3	1.6558	-0.2558	.0654
2	5.3	4.1	3.8626	1.4374	2.0661
3	4.8	5.6	5.7016	-0.9016	.8129
4	6.5	7.2	7.6632	-1.1632	1.3530
5	11.0	9.2	10.1152	0.8848	.7829
					5,0803

Figure 3. The Residuals and Sum of Squares from Figure 2.

Rank Regression

For the set of data given in Figure 1 the linear regression model appears to be satisfactory, so no further analysis would usually be required. However, merely for the sake of illustration, rank regression

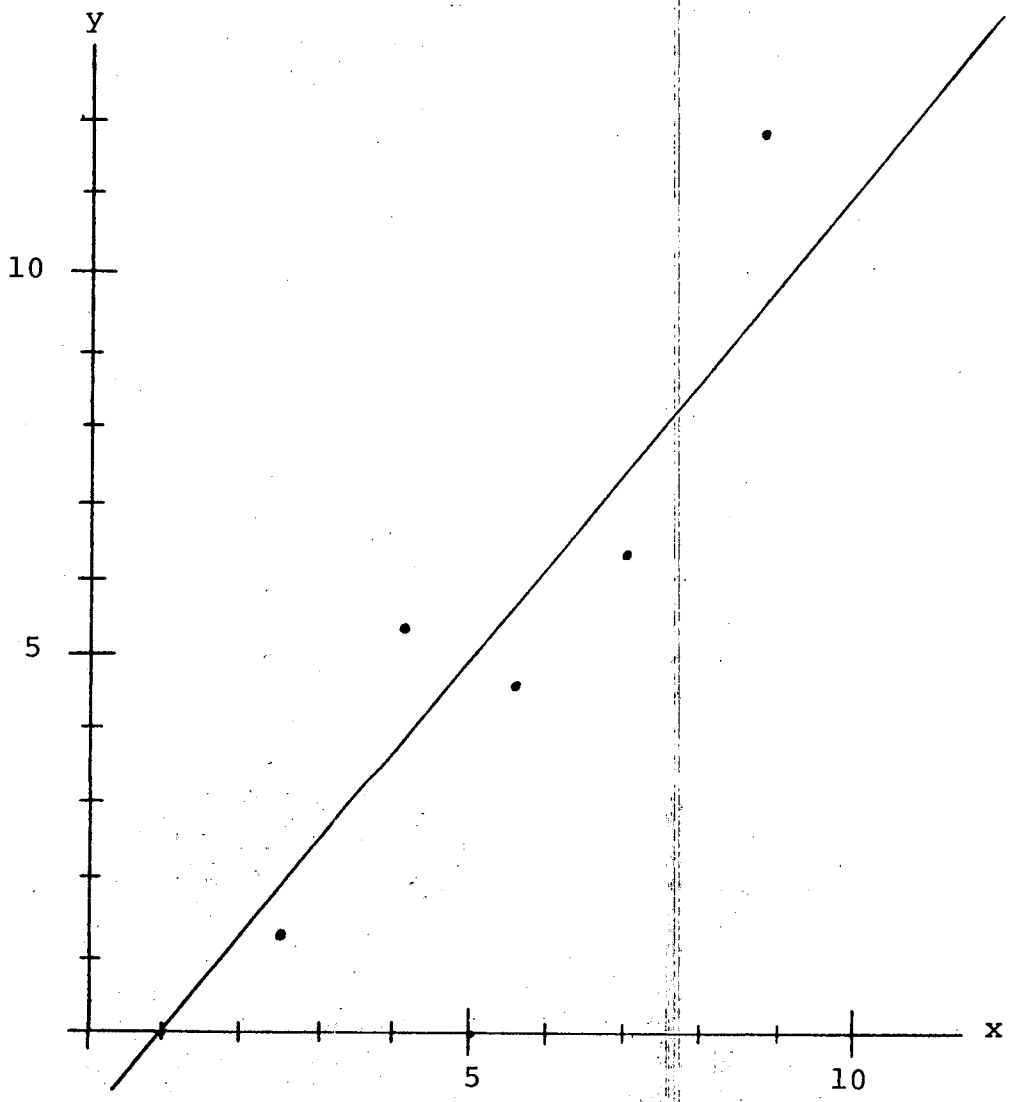


Figure 2. A Graph of the Data in Figure 1, and the Least Squares Regression Line.

is used on the same set of data and comparisons are made with regression on the raw data.

Rank regression involves simply the usual regression methods applied to the ranks of the data rather than to the data themselves. The smallest observation has rank 1, the second smallest has rank 2, and so on to the largest of n observations which has rank n. In case several observations in a group are all exactly equal to each other (tied), the rank assigned to each is the average of the ranks that would have been assigned to them had they not been tied. This is called the average ranks method of handling ties. In rank regression each variable is ranked by itself; that is, the observations on Y are ranked separately from the observations on X_1, which are in turn ranked separately from the observations on X_2, and so on. The ranks of the data in Figure 1 are given in Figure 4. Note that the ranks r_y of the Y's are obtained independently of the ranks r_x of the X's.

The least squares equation on the ranks is obtained using Equations (1) and (2) just as on the original data.

$$\hat{b}_r = \frac{\Sigma r_x r_y - (\Sigma r_x)(\Sigma r_y)/n}{\Sigma r_x^2 - (\Sigma r_x)^2/n} \qquad (3)$$

$$= \frac{54 - (15)(15)/5}{55 - (15)^2/5} = \frac{9}{10} = .9$$

$$\hat{a}_r = \frac{1}{n}(\Sigma r_y - b\Sigma r_x) \qquad (4)$$

$$= \frac{1}{5}(15 - (.9)(15)) = .3$$

Therefore the least squares equation on the ranks is given by

$$\hat{r}_y = \hat{a}_r + \hat{b}_r r_x \qquad (5)$$

$$= .3 + .9 r_x$$

Obs. Pair	Y	X	r_y	r_x	r_x^2	$r_x r_y$
1	1.4	2.3	1	1	1	1
2	5.3	4.1	3	2	4	6
3	4.8	5.6	2	3	9	6
4	6.5	7.2	4	4	16	16
5	11.0	9.2	5	5	25	25
Total			15	15	55	54

Figure 4. Worksheet for Least Squares on the Ranks

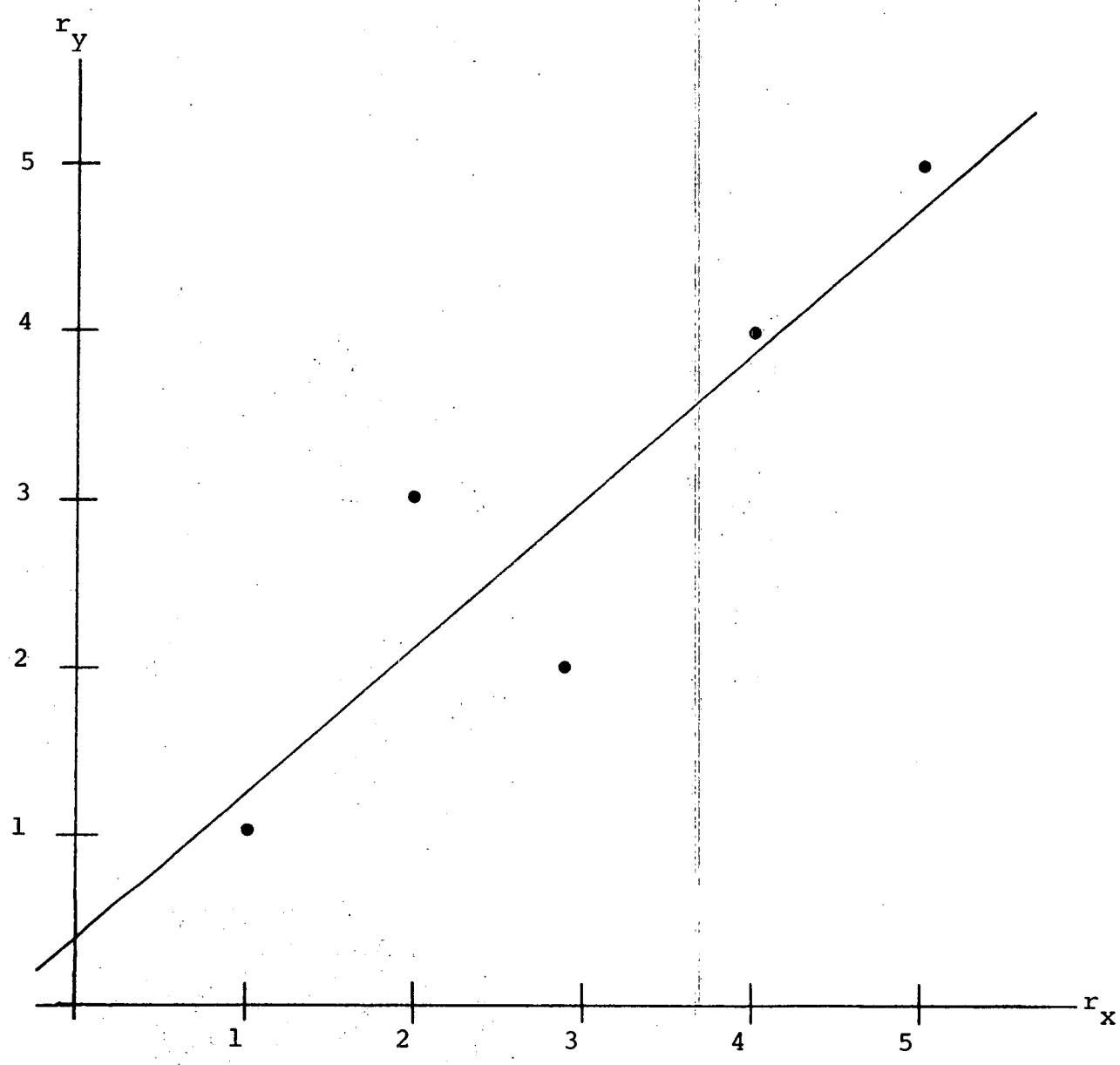


Figure 5. A Graph of the Ranks from Figure 4 and the Least Squares Line on the Ranks.

A graph of the ranks of the data, and the least squares straight line computed for the ranks is given in Figure 5. Again note that the linear model fits the ranks fairly well, just as the linear model fit the data fairly well in Figure 2. This is a typical relationship between rank regression and ordinary regression. If the simple linear regression model fits the data well, it usually works well on the ranks also. However, rank regression is useful in situations where simple linear regression does not work satisfactorily with the data. This point will be illustrated in a later example, but first the residuals from the rank regression procedure will be computed.

Converting Predicted Ranks to Predicted Values

There are two types of residuals from rank regression. One type of residual is the difference between the predicted ranks of Y and the actual rank of Y, which is obtained in the same way $\hat{Y}$ was obtained in ordinary regression, as in Figure 2, but using ranks r_y and r_x instead of Y and X, and using the equation for ranks Equation (5) instead of the least squares line for the data. These are called rank residuals and are not useful because they convey no information on how well the data are being fitted, only information on how well the ranks are being fitted.

To see how well the data are being fitted, the predicted rank $\hat{r}_y$ of each observation Y is obtained from Equation (5). These predicted ranks are converted to predicted values $\hat{Y}$ of Y by comparing the predicted ranks with the actual ranks of the five observations on Y, and obtaining predicted values of Y on the basis of this comparison, using interpolation if necessary.

Obs. Pair	r_y	r_x	$\hat{r}_y$	$\hat{Y}$	Y	Residual	$(Y-\hat{Y})^2$
1	1	1	1.2	2.08	1.4	.68	.4624
2	3	2	2.1	4.85	5.3	-.45	.2025
3	2	3	3.0	5.30	4.8	.50	.2500
4	4	4	3.9	6.38	6.5	-.12	.0144
5	5	5	4.8	10.10	11.0	-.90	.8100
Total							SS = 1.7393

Figure 6. Worksheet for Finding Residuals from Rank Regression

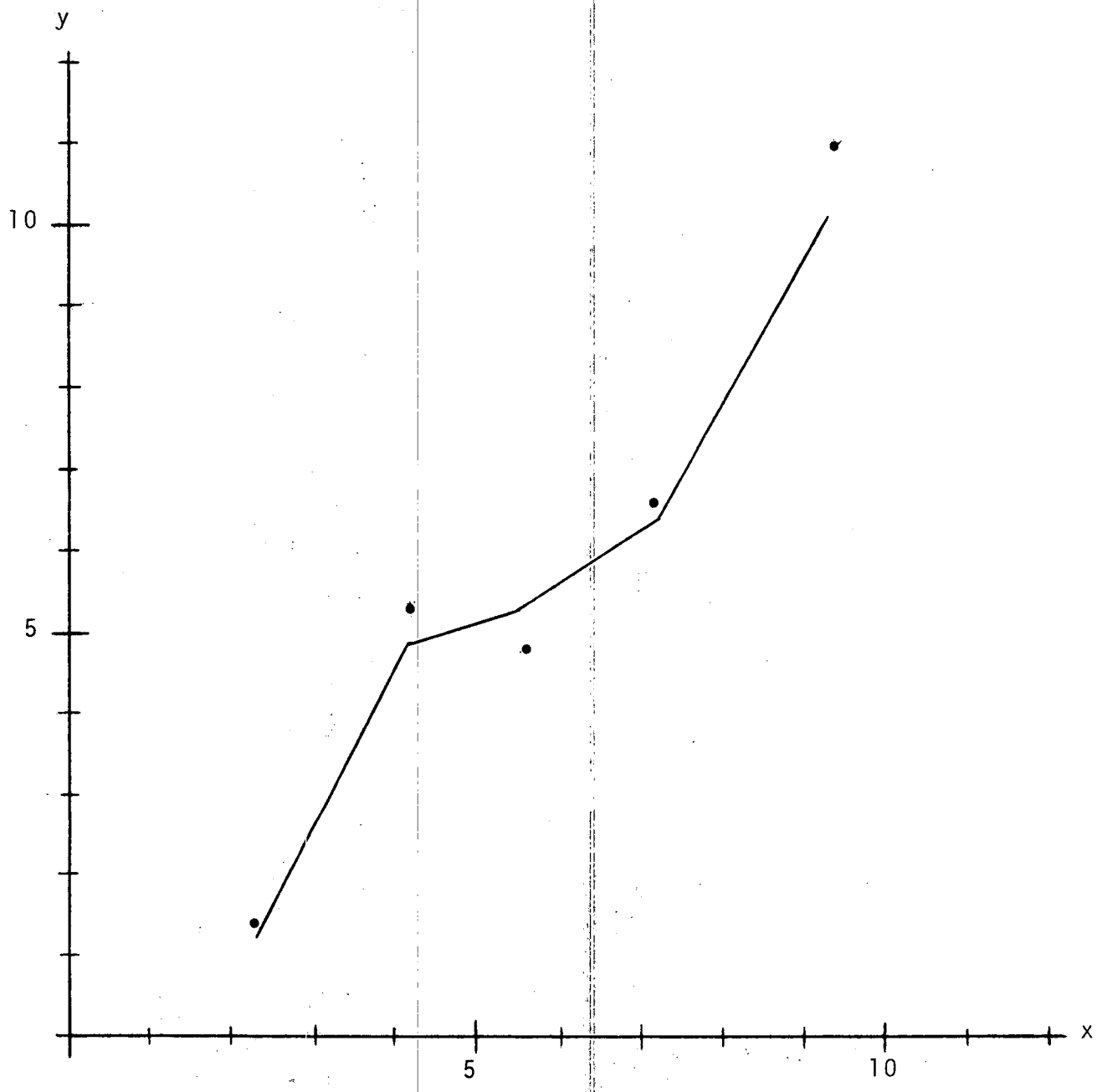


Figure 7. A graph of the Data in Figure 1, and the Rank Regression Curve

For example, to find the $\hat{Y}$ corresponding to the first pair of observations $(y_1, x_1) = (1.4, 2.3)$, enter the rank of x_1, $r_x = 1$, into Equation (5) to get the predicted rank of y_1;

$$\hat{r}_y = .3 + .9\, r_x$$

$$= .3 + (.9)(1)$$

$$= 1.2$$

The predicted Y corresponding to a predicted rank 1.2 is found by interpolating between the actual observed Y with rank 1, Y = 1.4, and the actual observed Y with rank 2, Y = 4.8;

$$\hat{Y} = 1.4 + \frac{1.2 - 1.0}{2.0 - 1.0}(4.8 - 1.4) = 2.08$$

This predicted value $\hat{Y} = 2.08$ is compared with the observed value in the first pair Y = 1.4 to get the residual 0.68. A summary of the calculations, including the residuals and the sum of squares is given in Figure 6.

The Residuals Sum of Squares (SS)

Note that the residuals in Figure 6 tend to be smaller than the residuals found from the least squares fit to the original data, given in Figure 3. The sum of squares from Figure 6

$$SS = 1.7393 \qquad \text{(Rank Regression)}$$

is much smaller than the sum of squares from Figure 3.

$$SS = 5.0803 \qquad \text{(Regression on Data)}$$

The residuals sum of squares from Figure 3 is the smallest that can be obtained from a straight line fit to the data. But rank regression does not give a straight line fit to the data. To show this, a graph of all possible rank regression predictions is given in Figure 7. Note that the rank regression equation adapts itself to the points observed, but is steadily increasing as x increases. Rank regression equations are monotonically increasing or decreasing, and therefore work very well with data that tend to show a monotonic relationship, even though the relationship may be nonlinear.

Exercise 1

Use the following steps to find the predicted value $\hat{Y} = 4.85$, for the second pair of observations $(y_2, x_2) = (5.3, 4.1)$, as given in Figure 6 from rank regression. See Figure 4 for the original data.

1. Find the rank of $x_2 = 4.1$.

2. Substitute the rank of x_2 into Equation (5) to get a predicted rank for y_2. Compare with the value from Figure 6.

3. Find the two observed values of Y whose ranks straddle (just above and just below) the predicted rank for y_2.

4. Interpolate between the two observed values for Y from step 3, to get a predicted value for y_2 that corresponds to the predicted rank for y_2 from step 2. This predicted value should match the value $\hat{Y} = 4.85$ from Figure 6.

The Flexibility of Rank Regression for Fitting Monotonic Data

The ability of the rank regression curve to adapt to a set of points which exhibit a nonlinear, but monotonic relationship is shown more dramatically in Figure 9. In Figure 8 nine points are obtained from the equation $y = e^x$, with no error of measurement added. The basic premise here is that a good regression technique should work well if the conditions are ideal. Here the conditions for rank regression are ideal, since the relationship between X and Y is monotonic. Because of this monotonic relationship between X and Y, the ranks of X show an exactly linear relationship with the ranks of Y. That is, the smallest X is paired with the smallest Y, so rank 1 for X is paired with rank 1 for Y. The second smallest X is paired with the second smallest Y so rank 2 for X is plotted against rank 2 for Y, and so on for all of the ranks. The ranks for the data in Figure 8 are graphed in Figure 10. This results in a rank regression equation in Figure 9 which consists of a series of line segments connecting the observed points. The fit is excellent. The least squares straight line is shown also, to dramatize the limitations of that method.

An Example with Real Data

The first two examples both involve artificial data. The first example serves merely to introduce the methodology of simple linear regression and rank regression. The second example illustrates a monotonic relationship between X and Y, and shows the ability of rank regression to adapt to data of this type. A third example will now be presented. It involves real data, where the independent variable X represents chemical measurements obtained using a relatively inexpensive titration method, and the dependent variable Y represents corresponding measurements obtained by a more expensive extraction and weighing technique. Twenty samples were obtained, and each sample was thoroughly mixed just before being split and analyzed by both methods.

This set of data is presented and thoroughly analyzed by Daniel and Wood (1971). Other authors have used these data in their papers on new regression methods, so this set of data is now a classical standard on which

X:	-2	-1.5	-1	-0.5	0	0.5	1	1.5	2
$Y = e^X$:	.14	.22	.37	.61	1.00	1.65	2.72	4.48	7.39

Figure 8. Nine Values from the Function $Y = e^X$.

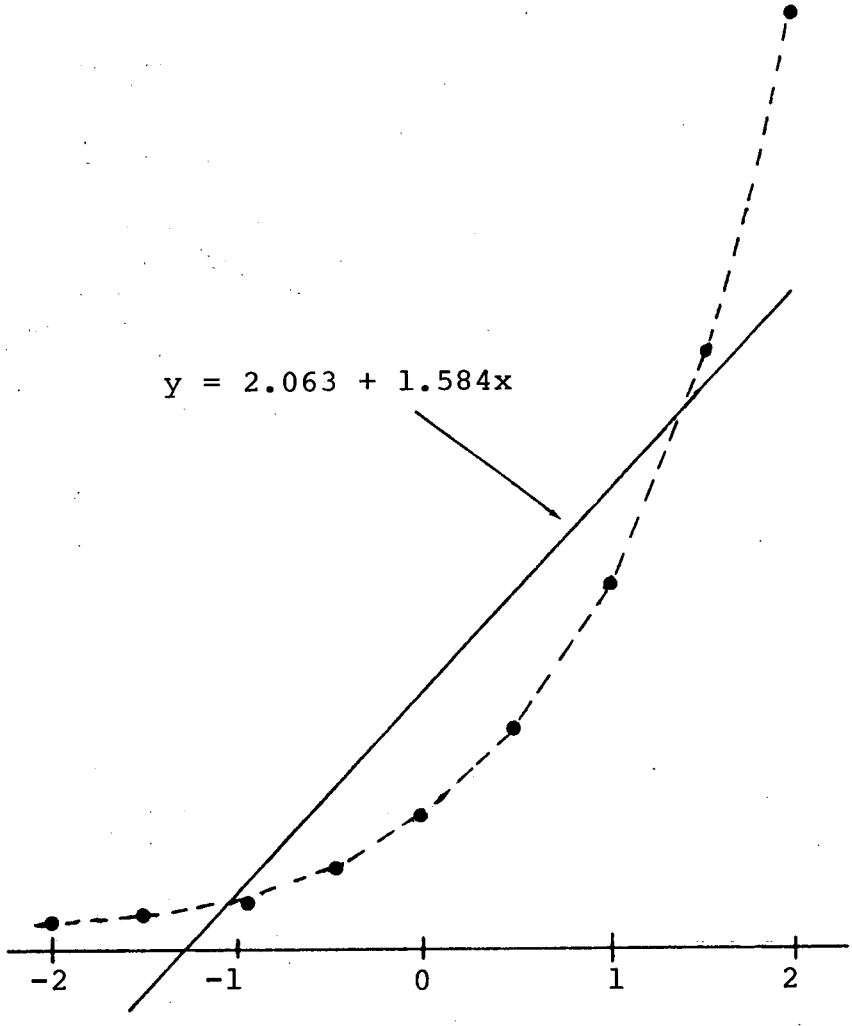


Figure 9. Rank Regression Curve and the Least Square Line for 9 Points from the Relationship $Y = e^X$.

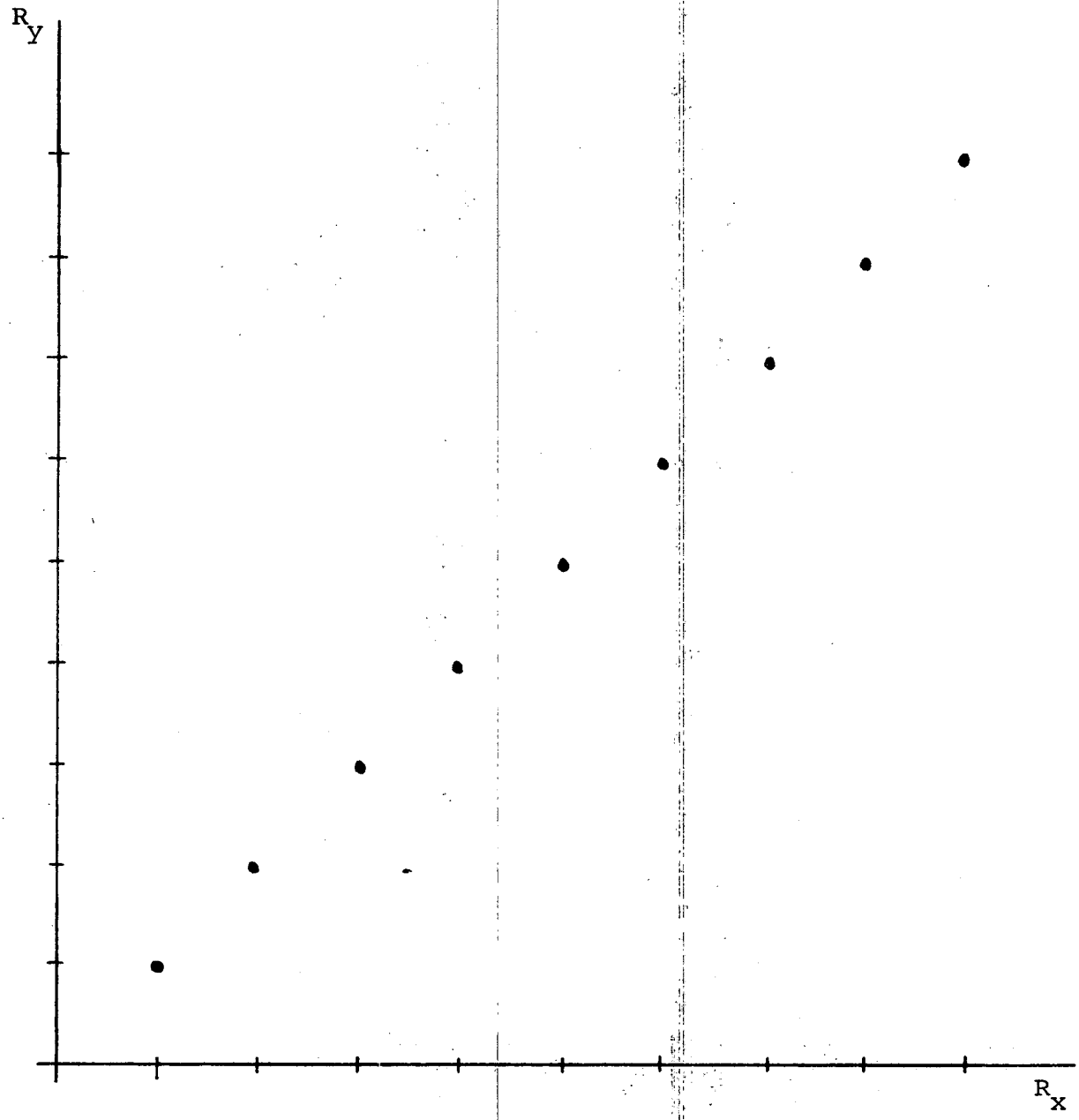


Figure 10. Graph of the Ranks of X and Y as Given in Figure 9.

new regression methods are tested. The data are somewhat unrealistic in that they follow a linear relationship very closely, with a correlation coefficient of .997. Most bivariate data sets encountered in applied work have a smaller correlation, are somewhat nonlinear in appearance, and contain occasional outliers which cannot be discarded because they represent legitimate measurements. Rank regression is more suited to the messy types of data, with nonlinearities and outliers, than to this type of data which is adequately explained by ordinary linear regression. But merely for the sake of illustration, rank regression is applied to this set of data and the results are compared with the fit from ordinary regression. The data and the residuals are given in Figure 11.

			Least Squares		Rank Regression	
Obs. No.	Y	X	$\hat{Y}$	Residual	$\hat{Y}$	Residual
1	76	123	75.02	.98	75.85	.15
2	70	109	70.51	- .51	70.98	- .98
3	55	62	55.40	- .40	56.11	-1.11
4	71	104	68.91	2.09	68.01	2.99
5	55	57	53.79	1.21	55.00	0.00
6	48	37	47.36	.64	48.18	- .18
7	50	44	49.61	.39	50.26	- .26
8	66	100	67.62	-1.62	66.04	- .04
9	41	16	40.60	.40	41.23	- .23
10	43	28	44.46	-1.46	43.51	- .51
11	82	138	79.84	2.16	81.75	.25
12	68	105	69.23	-1.23	69.99	-1.99
13	88	159	86.59	1.41	85.49	2.51
14	58	75	59.58	-1.58	58.25	- .25
15	64	88	63.76	.24	64.06	- .06
16	88	164	88.20	- .20	88.00	0.00
17	89	169	89.81	- .81	88.95	.05
18	88	167	89.17	-1.17	88.00	0.00
19	84	149	83.38	.62	83.89	.11
20	88	167	89.17	-1.17	88.00	0.00
				SS = 27.23		SS = 21.94

Figure 11. A Comparison of Least Squares Linear Regression on the Data with Regression on the Ranks, Using Data from Daniel and Wood (1971).

Comparing Rank Regression With Ordinary Regression

A comparison of the residuals in Figure 11 shows that the residuals from the rank regression are smaller than the residuals from ordinary linear regression for 15 out of the 20 points. The sum of squares of the residuals is only 21.94 for rank regression, compared with 27.23 for ordinary linear regression. The point to be made with this example is tht rank regression works well even with data that follow a close linear pattern.

The analysis of the data in Figure 11 was performed using the regression program described in Iman et al (1980). The slight difference in these results and the results reported in Iman and Conover (1979) is due to a difference in the method of handling ties.

Multiple Regression

The regression examples with one independent variable are useful for illustrating the principles behind ordinary regression and rank regression. When the number of independent variables is two or more these same principles apply but they become very difficult to illustrate. With two independent variables the least squares method on the data is used to fit a plane to data in the three dimensional space spanned by Y, X_1 and X_2. The rank regression method uses the least squares method to fit a plane to the ranks of the data, in the three dimensional space spanned by the ranks of Y, the ranks of X_1, and the ranks of X_2. When this plane is translated back to the three dimensional space spanned by Y, X_1 and X_2, the result is a series of connected mini planes that adapt to the data, in a monotonic manner, just as the series of line segments adapted to the data in the case of one independent variable. The extension to include more than two independent variables is simple in concept, but impossible to visualize because the discussion involves hyperplanes in many dimensional space.

An Example of Multiple Regression

An example is now presented which illustrates the results of ordinary multiple regression and multiple regression on the ranks. The data given in Figure 12 are from Brownlee (1965), and have become somewhat of a standard set of data for use in comparing new regression methods with old methods. They follow a linear regression pattern closely, with $R^2 = .914$. The measure of fit for a regression model is usually R^2, which states the proportion of variability of Y that is explained by the regression model. An R^2 of .914 means that 91.4% of the variation in Y is explained by regression on the variables X_1, X_2 and X_3. This figure is much closer to 100% than the R^2 values normally encountered in applied work.

The data in Figure 12 represent 21 successive days of operation of a plant oxidizing ammonia to nitric acid. The variables in Figure 12 are as follows:

Y = 10 times the percentage of the ingoing ammonia that is lost as unabsorbed nitric oxides; it is an indirect measure of the yield of nitric acid.

X_1 = the flow of air to the plant

Obs. No.	Y	X_1	X_2	X_3
1	42	80	27	89
2	37	80	27	88
3	37	75	25	90
4	28	62	24	87
5	18	62	22	87
6	18	62	23	87
7	19	62	24	93
8	20	62	24	93
9	15	58	23	87
10	14	58	18	80
11	14	58	18	89
12	13	58	17	88
13	11	58	18	82
14	12	58	19	93
15	8	50	18	89
16	7	50	18	86
17	8	50	19	72
18	8	50	19	79
19	9	50	20	80
20	15	56	20	82
21	15	70	20	91

Figure 12. Multivariate Data From Brownlee (1965).

Obs. No.	Y	Least Squares $\hat{Y}$	Least Squares Residual	Rank Regression $\hat{Y}$	Rank Regression Residual
1	42	38.77	3.23	41.64	0.36
2	37	38.92	-1.92	41.83	-4.83
3	37	32.44	4.56	37.00	0.00
4	28	22.30	5.70	19.01	8.99
5	18	19.71	-1.71	18.00	0.00
6	18	21.01	-3.01	18.36	-0.36
7	19	21.39	-2.39	18.85	0.15
8	20	21.39	-1.39	18.85	1.15
9	15	18.14	-3.14	15.00	0.00
10	14	12.73	1.27	12.47	1.53
11	14	11.36	2.64	12.21	1.79
12	13	10.22	2.78	11.11	1.89
13	11	12.43	-1.43	12.43	-1.43
14	12	12.05	-0.05	13.43	-1.43
15	8	5.64	2.36	8.00	0.00
16	7	6.09	.91	8.00	-1.00
17	8	9.52	-1.52	8.87	-0.87
18	8	8.46	-0.46	8.86	-0.86
19	9	9.60	-0.60	10.70	-1.70
20	15	13.59	1.41	12.80	2.20
21	15	22.24	-7.24	18.66	-3.66
			SS = 178.83		SS = 142.63

Figure 13. Predicted Values and Residuals Using Least Squares Regression and Rank Regression, on the Data from Figure 12, Using the Variables X_1, X_2 and X_3.

X_2 = The temperature of the cooling water entering the countercurrent nitric oxide absorption tower

X_3 = the concentration of nitric acid in the absorbing liquid

Ordinary Multiple Regression Illustrated

The model used to fit the data is

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 \qquad (6)$$

The method of least squares is used to find the values of β_0, β_1, β_2 and β_3 that minimize the residual sum of squares. The equations for finding the least squares coefficients are very complex and are not presented. The actual values for the coefficients appear as part of the computer printout, and are given in Equation (7).

$$y = -39.92 + .7156x_1 + 1.2953x_2 - .1521x_3 \qquad (7)$$

In order to evaluate the goodness of fit of this model, values for X_1, X_2 and X_3 are substituted into Equation (7) to get a predicted value $\hat{Y}$ for Y. For example, in observation number 1 in Figure 12, $X_1 = 80$, $X_2 = 27$ and $X_3 = 89$. These are substituted into Equation (7) as follows,

$$\hat{Y} = -39.92 + (.7156)(80) + (1.2953)(27) - (.1521)(89) = 38.76$$

to get 38.76 as the predicted value for Y. This is compared with the observed value for Y, 42, given in Figure 12. The residual (42 - 38.76) = 3.24 agrees with the value given in Figure 13, except for differences caused by rounding off the coefficients in Equation (7). The sum of squares of residuals is a measure of the goodness of fit of the model. In this case it is given by

$$SS = \Sigma(Y - \hat{Y})^2$$

$$= 178.83$$

as shown in Figure 13. This is the smallest value of SS possible using the model given by Equation (6).

Exercise 2

Substitute the values for X_1, X_2 and X_3, given in Figure 12 under observation number 2, into Equation (7) to verify that the predicted value in Figure 13 is indeed correct.

Rank Multiple Regression Illustrated

If the model in Equation (6) is used on the Ranks instead of the data it becomes

$$r_y = \beta_0 + \beta_1 r_{x_1} + \beta_2 r_{x_2} + \beta_3 r_{x_3} \qquad (8)$$

where r_y represents the ranks of the observations on Y, r_{x_1} represents the ranks of the observations on X_1, and so on, just as in the simpler examples given earlier. The ranks for the data in Figure 13 are given in Figure 14. The least squares method is used to find the values of the coefficients β_0 through β_3, which appear as part of the computer output. For the data in Figure 12 the coefficients are

$$r_y = -0.31 + .6650 r_{x_1} + .3859 r_{x_2} - .0226 r_{x_3} \qquad (9)$$

The predicted rank for each observation is obtained by substituting the respective ranks of the independent variables into Equation (9). As an example the ranks of X_1, X_2 and X_3 for observation number 1 are given in Figure 14 as $r_{x_1} = 20.5$, $r_{x_2} = 20.5$ and $r_{x_3} = 15$ respectively. These are substituted into Equation (9) to get

$$\hat{r}_y = -0.31 + (.6650)(20.5) + (.3859)(20.5) - (.0226)(15)$$

$$= 20.89$$

in agreement with the number given in Figure 14 for observation number 1. To convert this predicted rank, $\hat{r}_y = 20.89$, to a predicted value for Y the two values of Y whose ranks straddle 20.89 are found from Figure 14. These are Y = 42, whose rank is 21, and Y = 37, whose rank is 19.5. Interpolation gives the predicted value of Y,

$$\hat{Y} = 37 + \frac{20.89 - 19.5}{21 - 19.5}(42 - 37)$$

$$= 41.63$$

which agrees with the value from Figure 13, except for differences caused by using rounded-off values of the coefficients in Equation (9).

Exercise 3

Obtain the predicted rank in observation number 4 by substituting the ranks from Figure 14 into Equation (9). See if this agrees with the predicted rank given in Figure 14.

Exercise 4

Use the predicted rank $\hat{r}_y$ for observation number 4 from Figure 14 to obtain a predicted value for Y. See if this agrees with the number given in Figure 13.

Comparing Ordinary Regression and Rank Regression

A measure of the goodness of fit of the rank regression model is obtained by comparing each predicted value $\hat{Y}$ with the corresponding observed value Y. The sum of squares of the residuals is

$$SS = \Sigma(Y - \hat{Y})^2$$

$$= 142.63$$

which is less than the value 178.83 obtained using ordinary regression. Although ordinary regression finds the best fitting hyperplane, rank regression is not restricted to working with a single hyperplane. Actually a series of connected hyperplanes is obtained using rank regression, so the model can adjust to the data with some degree of flexibility, within the constraint of being monotone in each of its variables. Any disagreement between the residuals in Figure 13 and the residuals reported in Iman and Conover (1979) is due to a difference in the method of handling ties.

Sensitivity Analysis

It is apparent from the relative size of the coefficients in Equations (7) and (9) that the output is more sensitive to changes in some independent variables than to changes in others. For example, X_1 ranges from 50 to 80, a change of 30 units. The coefficient of X_1 in Equation (7) is .7156, so the maximum change in Y due to changes in X_1 is (30)(.7156) = 21.5 units. On the other hand X_3 ranges from 72 to 93, a distance of 21 units, and has a coefficient of -.1521 in Equation (7). The maximum change in Y due to changes in X_3 is only 3.2 units, less than one-sixth of the total influence of X_1. The situation is further complicated by the fact that X_1 and X_3 have a positive correlation coefficient of r_{13} = .500. This suggests that if X_3 were dropped from the regression model, the coefficient of X_1 might increase somewhat to account for some of the variability in Y that was formerly accounted for by X_3. Thus X_3 might not be making a significant contribution in the model, and perhaps should be omitted. Statistical tests are available for aiding in the decision of whether or not to omit variables from the model.

A similar line of reasoning may be used on the least squares fit to the ranks, only here the analysis is simpler because all of the ranks have approximately the same range. Only the coefficients in Equation (9) need to be examined. The coefficient of r_{X_3} -.0226 is about one-thirtieth the size of the coefficient of r_{X_1}, .6650, again suggesting that the rank of Y is much less sensitive to changes in the rank of X_3 than to changes in the rank of X_1.

Obs. No.	Y	r_y	$\hat{r}_y$	X_1	r_{x_1}	X_2	r_{x_2}	X_3	r_{x_3}
1	42	21	20.89	80	20.5	27	20.5	89	15
2	37	19.5	20.95	80	20.5	27	20.5	88	12.5
3	37	19.5	19.27	75	19	25	19	90	17
4	28	18	16.01	62	15	24	17	87	9.5
5	18	14.5	14.47	62	15	22	13	87	9.5
6	18	14.5	15.04	62	15	23	14.5	87	9.5
7	19	16	15.77	62	15	24	17	93	20
8	20	17	15.77	62	15	24	17	93	20
9	15	12	11.39	58	9.5	23	14.5	87	9.5
10	14	9.5	7.47	58	9.5	18	4	80	3.5
11	14	9.5	7.21	58	9.5	18	4	89	15
12	13	8	6.11	58	9.5	17	1	88	12.5
13	11	6	7.43	58	9.5	18	4	82	5.5
14	12	7	8.64	58	9.5	19	8	83	20
15	8	3	2.89	50	3	18	4	89	15
16	7	1	3.07	50	3	18	4	86	7
17	8	3	4.75	50	3	19	8	72	1
18	8	3	4.73	50	3	19	8	79	2
19	9	5	5.85	50	3	20	11	80	3.5
20	15	12	7.80	56	6	20	11	82	5.5
21	15	12	15.50	70	18	20	11	91	18

Figure 14. The Ranks of the Data in Figure 13 and the Predicted Ranks from Equation (9).

On Deciding What Variables to Include in the Model

The inclusion of many variables in the regression model may result in "overfitting" the data. That is, the effect of having many variables is to force the regression surface into wildly erratic patterns just so it will pass closer to the observed points and have smaller residuals. Some systematic method is needed to assist in deciding whether variables should be included or excluded from the analysis. With such a tool to aid in the decision making, one may consider other variables related to X_1, X_2 and X_3, such as X_1^2 or X_1X_2, because it is possible that these other variables are more useful than the simple ones that have been considered so far. Three of these decision-assisting tools will be introduced and compared. But first the concept of partial correlation needs to be explained.

Simple Correlation

The strength of a simple linear relationship between two variables in usually measured with r, called Pearson's product moment correlation coefficient, or simply, the correlation coefficient for short. The correlation coefficient between X and Y is given by

$$r = \frac{\Sigma(X_i - \overline{X})(Y_i - \overline{Y})}{\sqrt{\Sigma(X_i - \overline{X})^2 \, \Sigma(Y_i - \overline{Y})^2}} \qquad (10)$$

for paired observations observations $(X_1, Y_1), \ldots, (X_n, Y_n)$, where $\overline{X}$ and $\overline{Y}$ are the sample means. The statistic r may be used to test the hypothesis of no correlation, but only if the variables have a particular distribution called bivariate normal distribution. This condition is often assumed, but rarely met, in practice.

When more than two variables are involved subscripts are used to show which two variables are being correlated. Thus r_{12} refers to the correlation between X_1 and X_2,

$$r_{12} = \frac{\Sigma(X_{1i} - \overline{X}_1)(X_{2i} - \overline{X}_2)}{\sqrt{\Sigma\,(X_{1i} - \overline{X}_1)^2 \, \Sigma(X_{2i} - \overline{X}_2)^2}} \qquad (11)$$

while r_{y1} refers to the correlation between Y and X_1.

Rank Correlation

The strength of a monotonic relationship between two variables, as opposed to a linear relationship, is usually measured using r, but computed on the ranks of the variables instead of the variables themselves. It is customary to use either r_s or ρ (rho) to denote the rank correlation coeffi-

cient, sometimes called Spearman's rho. We will try to minimize problems with subscripts by using ρ, such as ρ_{12}

$$\rho_{12} = \frac{\Sigma\left(r_{x_{1i}} - \frac{n+1}{2}\right)\left(r_{x_{2i}} - \frac{n+1}{2}\right)}{\sqrt{\Sigma\left(r_{x_{1i}} - \frac{n+1}{2}\right)^2 \Sigma\left(r_{x_{2i}} - \frac{n+1}{2}\right)^2}} \tag{12}$$

to denote the correlation between the ranks r_x of X_1 and the ranks r_x of X_2. All of the calculations and interpretation of results using the correlation on the data may be applied to the ranks of the data as well. The regression program described by Iman, et al (1980) handles all of the calculations on the data or the ranks of the data, the calculations just described and the calculations in the following pages. Spearman's rho may be used to test the hypothesis of independence, without requiring any distributional assumptions. Special tables may be found in Conover (1980).

Partial Correlation

Sometimes the apparent correlation between two variables may be due in part to the indirect influence of a third variable on both of the other variables. For example, the weekly average municipal bond yield X_1 may appear to be correlated with the average utility bond yield X_2 for the same week. Yet both may be heavily influenced by the average interest rate charged by the Federal Reserve, X_3, for that week. How can the influence of the variable X_3 be removed from the relationship between X_1 and X_2?

One way to do this is to use a simple linear regression of each variable X_1 and X_2 separately on X_3. The least squares method is used to fit coefficients to

$$\hat{X}_1 = \beta_0 + \beta_1 X_3$$

to get the residuals $(X_1 - \hat{X}_1)$. In a similar manner the residuals $(X_2 - \hat{X}_2)$ are also obtained by a regression of X_2 on X_3. Thus the linear influence of X_3 on both X_1 and X_2 is removed, and the correlation between the residuals $(X_1 - \hat{X}_1)$ and $(X_2 - \hat{X}_2)$ can be computed using Equation (10), but where r is computed using $(X_1 - \hat{X}_1)$ instead of X_1, and $(X_2 - \hat{X}_2)$ instead of X_2.

An Equation for Computing Partial Correlation

Such a correlation coefficient is called the partial correlation coefficient between X_1 and X_2, given X_3 and is denoted by $r_{12.3}$. An easy way to compute $r_{12.3}$ is with the equation

$$r_{12.3} = \frac{r_{12} - r_{13}\, r_{23}}{\sqrt{\left(1 - r_{13}^2\right)\left(1 - r_{23}^2\right)}} \tag{13}$$

where r_{12}, r_{13} and r_{23} are the simple linear correlation coefficients between X_1 and X_2, X_1 and X_3, and X_2 and X_3 respectively. The relationship given by Equation (13) obtains the same partial correlation coefficient as would be obtained by going through the process of correlating residuals described earlier.

Partial Rank Correlation

The use of rank regression, and the rank correlation coefficients ρ_{12}, ρ_{13}, and ρ_{23} gives $\rho_{12.3}$, which measures the strength of the linear association between the ranks of X_1 and the ranks of X_2, after the linear effect of the ranks of X_3 is removed from the ranks of X_1, and the ranks of X_2. The equation

$$\rho_{12.3} = \frac{\rho_{12} - \rho_{13}\rho_{23}}{\sqrt{\left(1 - \rho_{13}^2\right)\left(1 - \rho_{23}^2\right)}} \qquad (14)$$

may be used to find the partial rank correlation coefficient.

Partial Correlation Given Several Variables

The concept of partial correlation is easily extended to measure the strength of the linear relationship between two variables X_1 and X_2, after the linear effects of several variables, say X_3, X_4 and X_5, have been removed. One way of looking at this partial correlation is to compute the correlation on the residuals $(X_1 - \hat{X}_1)$ and $(X_2 - \hat{X}_2)$ as before, but where X_1 and X_2 are obtained from a linear regression on X_3, X_4, and X_5. That is, the model

$$\hat{X}_1 = \beta_0 + \beta_1 X_3 + \beta_2 X_4 + \beta_3 X_5$$

is fitted using least squares, and so is the model

$$\hat{X}_2 = \beta'_0 + \beta'_1 X_3 + \beta'_2 X_4 + \beta'_3 X_5$$

in order to remove the unwanted influence of X_3, X_4 and X_5 on the variables X_1 and X_2. The correlation between $(X_1 - \hat{X}_1)$ and $(X_2 - \hat{X}_2)$ is denoted by $r_{12.345}$. The corresponding treatment on the ranks is denoted by $\rho_{12.345}$.

Another way of finding $r_{12.345}$ is by the equation

$$r_{12.345} = \frac{r_{12.45} - r_{13.45}\, r_{23.45}}{\sqrt{(1 - r^2_{13.45})(1 - r^2_{23.45})}} \qquad (15)$$

which is analogous to Equation (13).

Three Multiple Regression Procedures

Correlation coefficients and partial correlation coefficients are used in several different ways to assist in deciding which variables to include in a regression model involving Y and several independent variables. Three such methods will be described.

1. The forward procedure. The regression model starts with no variables in it. Variables are then added, one at a time, to the regression model until a decision is reached that no new variables will contribute significantly to the improvement of the fit of the model to the data.

2. The backward procedure. The regression model starts with all of the variables in it. Then variables which are considered not making a significant contribution to the fit are dropped one by one until all insignificant variables have been removed from the model.

3. The stepwise procedure. The regression model starts with no variables in it. Variables are then added, one at a time, as in the forward procedure. Each time a variable is added, the backward procedure is used to delete all variables previously added to the model but which are now considered to be insignificant in the presence of the new variable. This combination of the forward and backward procedures is the most commonly used method of determining which set of variables belongs in the model.

The Variables Being Considered

These methods for assisting in deciding which variables to include enable many more variables to be considered than one would normally use in a model. In the backward procedure the number of variables must be less than the number of data points being fitted, but in the forward and stepwise procedures no such limitation exists. For the example introduced earlier, the original variables X_1, X_2 and X_3 will be considered, plus the squared terms X_1^2, X_2^2, X_3^2, and the cross products X_1X_2, X_1X_3, and X_2X_3. Thus nine independent variables will be considered in this regression model. Many other variables could be included, such as $1/X_1$, X_1/X_2, $\sqrt{X_3}$, $\ln X_2$, and so on. Any variables which may have a real physical interpretation in the situation being modeled should be considered. In our example we will consider only the nine first and second order terms, as mentioned. For convenience the variables, including Y, are numbered from 1 to 10, as shown in Figure 15.

The Forward Procedure

The forward procedure starts with a model that has no variables. The first variable chosen to be included in the model is the one that has the largest correlation with Y. The largest value of r for the data in Figure 12 and involving Y as one of the variables is the correlation coefficient between Y and X_1X_2

$$r_{18} = .9588$$

so X_1X_2 is the first variable considered for inclusion in the model.

A Test of Significance

At this point a test of significance is conducted to see if the inclusion of X_1X_2 in the model reduces the unexplained variation of Y a significant amount. The statistical test used is exact only if several assumptions regarding the distribution of Y are true. These assumptions are well explained in textbooks which present regression methods (see Draper and Smith, 1966), so they are not repeated here. In practice these assumptions are, at best, only approximately true, so this test and all subsequent statistical tests are only approximate. Still they serve as a useful objective method for assisting in making the decision as to whether or not a variable should be included. The test indicates (at $\alpha = .05$) that X_1X_2 should be included, so the forward procedure continues.

Using Partial Correlation in Forward Regression

The partial correlation coefficients of each variable with Y, given X_1X_2, are compared and the largest one is selected. Partial correlations are computed using Equation (13) in conjunction with the correlations given in Figure 15.

For example, the partial correlation of Y with X_1^2, given X_1X_2 is

$$r_{15.8} = \frac{r_{15} - r_{18}\, r_{58}}{\sqrt{(1 - r_{18}^2)(1 - r_{58}^2)}}$$

$$= \frac{.9251 - (.9588)(.9500)}{\sqrt{(1 - (.9588)^2)(1 - (.9500)^2)}}$$

$$= .1605$$

Exercise 5

Use Equation 13 and the correlation coefficients given in Figure 15 to find the partial correlation coefficient of Y and X_1, given X_1X_2.

Variable Number	Variable Name	Correlation With Y	Correlation With X_1X_2	Rank Correlation With Y	Rank Correlation With X_1X_2
1	Y	1.0000	.9588	1.0000	.9224
2	X_1	.9197	.9453	.9180	.9099
3	X_2	.8755	.9378	.8521	.9188
4	X_3	.3998	.4508	.4974	.4985
5	X_1^2	.9251	.9500	.8896	.9519
6	X_2^2	.8934	.9499	.8897	.9628
7	X_3^2	.3958	.4497	.4161	.4352
8	X_1X_2	.9588	1.0000	.9224	1.0000
9	X_1X_3	.8712	.9094	.7892	.8249
10	X_2X_3	.8554	.9239	.8276	.8862

Figure 15. Some Correlation Coefficients for the Data from Figure 12 and the Ranks from Figure 14.

Adding Another Variable to the Model

The largest partial correlation coefficient is found, and that variable is examined using the test of significance, to see if including it in the model reduces the variation of Y a significant amount. In this case the test shows nonsignificance, so no additional variables are included in the model.

The Forward Regression Model

Only the variable X_1X_2 is included in the model, of the nine variables examined, for the data in Figure 12. The least squares fit results in the model

$$\hat{Y} = -15.29 + .025315\ X_1X_2 \qquad (16)$$

The R^2 value is .919 for this model, indicating that 91.9% of the variation in Y is accounted for by a linear regression on X_1X_2. This is about the same as the $R^2 = .914$ obtained for the model in Equation (7) which used the variables X_1, X_2 and X_3. The inclusion of additional variables automatically increases the value of R^2, generally, a model with only one variable is preferred over a model with the same value of R^2 but with three variables.

Obtaining Predicted Values in Multiple Regression

To obtain a predicted value from Equation (16) the value of X_1 is multiplied by X_2, and the product is placed in the equation. For example, in observation number 1 in Figure 12, $X_1 = 80$ and $X_2 = 27$. Then Equation (16) becomes

$$\hat{Y} = -15.29 + .025315\ X_1X_2$$

$$= -15.29 + (.025315)(80)(27)$$

$$= 39.39$$

in agreement with the predicted value given in Figure 16. The sum of squares of residuals for this model is

$$SS = 116.846$$

which is considerably better (smaller) than the value 178.83 given in Figure 13 for the model with variables X_1, X_2 and X_3. This forward regression method has found a better model, with fewer variables, than the model examined earlier.

The Forward Rank Regression Procedure

Forward rank regression proceeds just as described for ordinary forward regression, except X_1, X_2 and X_3 are replaced by their ranks initially. Then terms corresponding to X_1^2 or X_1X_2 become the square of the rank of X_1, or the rank of X_1 times the rank of X_2.

The largest simple rank correlation is

$$\rho_{18} = .9224$$

so the rank analogue to X_1X_2 is selected for possible inclusion in the model. The test of significance indicates, at $\alpha = .05$, that this variable should be included in the model.

The largest partial correlation with the rank of Y, given the rank analogue of X_1X_2, involves the rank of the variable X_1

Obs. No.	Observed Value Y	Least Squares Model: X_1X_2 $\hat{Y}$	Residual	Rank Regression Model: X_1, X_2^2, X_1X_2 $\hat{Y}$	Residual
1	42	39.39	2.61	40.40	1.60
2	37	39.39	-2.39	40.40	-3.40
3	37	32.17	4.83	33.95	3.05
4	28	22.38	5.62	19.80	8.20
5	18	19.24	-1.24	16.59	1.41
6	18	20.81	-2.81	18.00	0.00
7	19	22.38	-3.38	19.80	-0.80
8	20	22.38	-2.38	19.80	0.20
9	15	18.48	-3.48	16.37	-1.37
10	14	11.14	2.86	12.70	1.30
11	14	11.14	2.86	12.70	1.30
12	13	9.67	3.33	13.44	-0.44
13	11	11.14	-0.14	12.70	-1.70
14	12	12.60	-0.60	13.10	-1.10
15	8	7.49	0.51	7.34	0.66
16	7	7.49	-0.49	7.34	-0.34
17	8	8.76	-0.76	8.00	0.00
18	8	8.76	-0.76	8.00	0.00
19	9	10.02	-1.02	11.81	-2.81
20	15	13.06	1.94	13.13	1.87
21	15	20.15	-5.15	17.06	-2.06
			SS = 116.846		SS = 119.176

Figure 16. Predicted Values and Residuals from the Forward Regression Models on the Data and on the Ranks.

$$\rho_{12.8} = \frac{\rho_{12} - \rho_{18}\,\rho_{28}}{\sqrt{(1 - \rho_{18}^2)(1 - \rho_{28}^2)}}$$

$$= \frac{.9180 - (.9224)(.9099)}{\sqrt{(1 - (.9224)^2)(1 - (.9099)^2)}}$$

$$= .4912$$

and this is large enough to be declared significant, using the same statistical test that is used for $r_{12.8}$. Therefore the variable X_1 is added to the model.

The next step consists of examining all partial correlations with the rank of Y, given the rank analogues to X_1 and X_1X_2. The largest partial rank correlation coefficient belongs to the rank analogue of X_2^2, and that term is included in the model after a statistical test determines its significance. No more terms are found to be significant. The least squares method is used to obtain the coefficients, with the result

$$\hat{r}_y = -2.11 + 1.1938\, r_{x_1} + .06227\, r_{x_2}^2 - .06666\, r_{x_1}\, r_{x_2} \qquad (17)$$

Using Rank Regression to Predict Values of Y

To find a predicted rank for Y, the corresponding ranks for X_1 and X_2 are substituted into Equation (17). For example, in observation number 1 of Figure 14 the ranks are $r_{x_1} = 20.5$ and $r_{x_2} = 20.5$. This gives

$$\hat{r}_y = -2.11 + 1.1938(20.5) + .06227(20.5)^2 - .06666(20.5)(20.5)$$

$$= 20.52$$

The two ranks which straddle the value 20.52 are, from Figure 14, 21 (Y = 42) and 19.5 (Y = 37). Interpolation gives

$$\hat{Y} = 37 + \frac{20.52 - 19.5}{21 - 19.5}(42 - 37)$$

$$= 40.40$$

which is in agreement with the value given in Figure 16.

Exercise 6

Use the ranks of X_1 and X_2 for observation number 9 in Figure 14 to find $\hat{r}_y$ from Equation 17. (The computer printout gives $\hat{r}_y = 13.14$.)

Exercise 7

Use interpolation in Figure 14 to obtain a predicted value for Y, given $\hat{r}_y = 13.14$ for observation number 9, and see if your answer agrees with the predicted value given in Figure 16.

Comparing Rank Regression with Ordinary Regression

The sum of squares of residuals for the rank regression model given by Equation (17) is

$$SS = 119.176$$

which is about the same as the value found using forward regression on the data. Of the two models found using forward regression, the model based on the data is preferred, because it has about the same value for SS as the model based on ranks, and it has only one variable as opposed to three variables for the model based on ranks.

Backward Regression

The backward regression procedure begins with the model containing all of the variables, fitted using the least squares method. Then the partial correlation coefficients of each variable with Y, given all of the other variables in the model, are compared to see which one is the smallest. The variable corresponding to this smallest partial rank correlation coefficient is then tested to see if its presence in the model contributes significantly in accounting for the variation in Y. If the test is significant the variable remains in the model and the procedure is finished. If the variable does not test as significant, it is dropped from the model, and the entire procedure is repeated for the remaining variables. At each stage the partial correlations involve only the variables remaining in the model. This backward regression procedure has a tendency to include more variables in the final model than if forward regression is used.

A Useful Procedure for Finding Partial Correlation Coefficients

At this point a very useful technique will be introduced and illustrated. The method for finding partial correlation coefficients by building step by step from the simple correlation coefficients, as described earlier, works well for forward regression. This is an awkward way of handling backward regression however, since backward regression starts with the full model, all variables included, and requires knowledge of the partial correlation coefficients given all but one of the independent variables. There is a very simple way to do this.

Let R be the correlation matrix; that is, the matrix containing the simple pairwise correlation coefficients where row i, column j contains r_{ij}.

Let R^{-1} be the inverse matrix of R, which is the matrix that can be multiplied by R to get the identity matrix I. Many software packages exist for finding R^{-1}. Denote the element in row i, column j of R^{-1} by b_{ij}. The partial correlation of variable i with variable j, given all of the other variables represented in the correlation matrix, is obtained very simply from R^{-1} as

$$r_{ij.(\text{all others})} = - \frac{b_{ij}}{\sqrt{b_{ii}\, b_{jj}}} \qquad (18)$$

When a variable is eliminated from the model because it is not significant, the row and column in R, which represent correlations involving that variable, are removed from R to obtain a new and smaller correlation matrix R_1. The inverse R_1^{-1} of R_1 is found and the new partial correlation coefficients of each variable with Y, given the variables still remaining in the model, are found from R_1^{-1} using Equation (18).

An Illustration of the Procedure

This procedure for finding partial correlation coefficients could be illustrated using the full model, with Y and nine independent variables. This would involve a correlation matrix R with 10 rows and 10 columns, and its inverse R^{-1} also with 10 rows and 10 columns. Such an example is so large that it may confuse the illustration of the procedure, so the procedure will be illustrated using only Y and the original independent variables X_1, X_2 and X_3.

For the data in Figure 12 the correlation matrix is given by

$$R = \begin{array}{c} \begin{matrix} X_1 & X_2 & X_3 & Y \end{matrix} \\ \begin{bmatrix} 1.0000 & .7819 & .5001 & .9197 \\ .7819 & 1.0000 & .3909 & .8755 \\ .5001 & .3909 & 1.0000 & .3998 \\ .9197 & .8755 & .3998 & 1.0000 \end{bmatrix} \end{array}$$

where row 4 (and column 4) contains the correlation coefficients of Y with X_1, X_2 and X_3 and itself, respectively.

The inverse matrix is found from a computer program as

$$R^{-1} = \begin{array}{c} \begin{matrix} X_1 & X_2 & X_3 & Y \end{matrix} \\ \begin{bmatrix} 7.7241 & .9922 & -1.2655 & -7.4665 \\ .9922 & 4.4469 & -.3727 & -4.6568 \\ -1.2655 & -.3727 & 1.4078 & .9274 \\ -7.4665 & -4.6568 & .9274 & 11.5732 \end{bmatrix} \end{array}$$

Note that both R and R-1 are symmetric matrices.

The partial correlation coefficients are found using Equation (18). Only the partial correlations involving Y are of interest, since these represent the amount of influence each independent variable has on Y, when the linear influence of the other variables is removed. These are

$$\text{Y and } X_1: \quad r_{41.23} = \frac{-b_{41}}{\sqrt{b_{44}\, b_{11}}} = \frac{7.4665}{\sqrt{(7.7241)(11.5732)}} = .7897$$

$$\text{Y and } X_2: \quad r_{42.13} = \frac{-b_{42}}{\sqrt{b_{44}\, b_{22}}} = \frac{4.6568}{\sqrt{(4.4469)(11.5732)}} = .6491$$

$$\text{Y and } X_3: \quad r_{43.12} = \frac{-b_{43}}{\sqrt{b_{44}\, b_{33}}} = \frac{-.9274}{\sqrt{(1.4078)(11.5732)}} = -.2298$$

In a backward regression procedure the variable X_3 would be selected for testing because it has the lowest partial correlation coefficient (in absolute value). If the test showed that X_3's contribution to the model were insignificant it would be dropped from the model. Row 3 and column 3 would be removed from the correlation matrix R, the new inverse matrix found, and the new partial correlation coefficients found with X_3 eliminated from all further consideration. This procedure would be repeated until all variables remaining in the model are significant.

Measuring the Importance of Variables

A direct interpretation of the partial correlation coefficients found from the full model is very useful. The three variables X_1, X_2 and X_3 may be ranked in order of importance on the basis of their partial correlation coefficients. Variable X_1 is the most important because its partial correlation coefficient, $r_{41.23} = .7897$, is the largest in absolute value. The variable X_2 is the second most important variable and X_3 is the least important variable. Although a rough analysis was made when this model was introduced in Equation (7), and resulted in these same conclusions, this method of analysis removes from consideration the indirect influence of the other variables, and thus isolates the effect of each variable on Y in a more precise manner.

The Results Using Backward Regression

The backward regression procedure on all nine independent variables removed the variable X_2^2 in its first step. Subsequent steps removed, one by one, the variables X_2X_3, X_1, X_3, X_1X_3, X_3^2, X_1^2 and X_2, leaving X_1X_2 as the only significant variable. Thus the backward regression procedure ended up with the same model as was obtained from the forward regression procedure. This is merely a coincidence that happened to occur in this case. The predicted values, the residuals, and the residual sum of squares are thus given in Figure 16. This lends further support to the conclusion that the model involving only X_1X_2 is a good model for the data.

Backward Rank Regression

Backward regression on the ranks follows the same steps as backward regression on the data. The only difference is that the ranks of Y, X_1, X_2 and X_3 are substituted for the actual values of those variables. The partial rank correlations are found by inverting the matrix of simple rank correlations and using

$$\rho_{ij.(\text{all others})} = \frac{-b_{ij}}{\sqrt{b_{ii}\, b_{jj}}} \qquad (19)$$

in a manner entirely analogous to the procedure described previously.

The backward regression procedure on the ranks given in Figure 14 results in the elimination of variable X_2 on the first step. Note that the analysis on the actual data differed in this respect, because X_2^2 was removed on the first step, and the variable X_2 was the last one to be removed.

Bear in mind that regression on ranks is based on monotonic relationships, while regression on the data is based on linear relationships, so removal of X_2 using rank regression is not entirely different than the removal of X_2^2 using ordinary regression.

The Model from Backward Rank Regression

Subsequent steps in the backward rank regression procedure removed, one at a time, the variables X_1X_3, X_2X_3, X_1X_2, X_3, and X_3^2. This left the variables X_1, X_1^2 and X_2^2 in the model. Several interesting things happened. The variable X_1X_2 was removed on the fourth step. This is one of the three variables which remained in the model using forward regression on the ranks, and the only variable which remained in the model using both forward regression and backward regression on the actual data. Also, the variable X_1^2 remains in the model for backward regression on the ranks. This is the first time this variable has appeared in a regression model. It is interesting to see what effect the replacement of X_1X_2 with X_1^2 has on the rank regression model, as this represents the only difference between the forward regression model on the ranks and the backward regression model on the ranks. The final model is given as

$$\hat{r}_y = -2.02 + 1.41\, r_{x_1} - .04410\, r^2_{x_1} + .02781\, r^2_{x_2} \qquad (20)$$

The predicted values of Y and the residuals appear in Figure 17.

A Comparison of the Several Models

The model given by Equation (20) appears to be the best model obtained so far, as measured by the sizes of the residuals given in Figure

17. When compared with forward rank regression, the backward rank regression model resulted in smaller residuals for 10 of the 21 points, larger residuals for 9 of the 21 points, and ties in 2 cases. However, the important measure of goodness of fit, the sum of squares of residuals, is reduced by about 23 percent, to

$$SS = 91.677$$

which is by far the smallest value for SS yet obtained. This is smaller than the value 116.846 obtained using the model using X_1X_2 on the actual data. One model has a smaller SS, while the other model has fewer variables. It is difficult to choose between these two models at this point.

Stepwise Regression

The stepwise regression procedure is the procedure that is used most often in obtaining a regression model. It begins as a forward regression procedure, but each time a variable is added to the model it becomes a backward regression procedure until all insignificant variables have been eliminated from the model. When a new variable is added to the model, the partial correlation coefficients consider this new variable in addition to the previous variables, and they may be different than the previous partial correlation coefficients. This is why a variable that was previously considered significant using forward regression may become insignificant after a new variable is added, in which case it would be dropped using backward regression. This does not preclude that variable from being added again at some later point, if it again becomes significant after additional variables have entered the model.

The application of the stepwise regression procedure to the data in Figure 12 and the ranks in Figure 14 is not interesting, because in those cases no variables are eliminated from the model as insignificant. Therefore the resulting models, predicted values, and residuals are the same as those obtained using forward regression, and given in Figure 16.

Discussion

Several regression procedures have been introduced and described for constructing a regression model involving several variables. Partial correlation is discussed as a tool for identifying important variables. The computations have been explained so that a better understanding of the powers and limitations of the various procedures can be obtained.

All of these procedures are merely aids in the decision making process. They should be considered in addition to expert advice, not instead of expert advice. Amateur statisticians often make the mistake of having either too little or too much faith in the methods of regression analysis. Professional statisticians often make the same mistake, but to a lesser extent. The better these regression methods are understood, the more likely it is that the results of a regression analysis will be given its proper weight in the final decision making process.

Obs. No.	Observed Value Y	Forward Rank Regression: X_1, X_2^2, X_1X_2 $\hat{Y}$	Residual	Backward Rank Regression: X_1, X_1^2, X_2^2 $\hat{Y}$	Residual
1	42	40.40	1.60	38.92	3.08
2	37	40.40	-3.40	38.92	-1.92
3	37	33.95	3.05	33.54	3.46
4	28	19.80	8.20	22.18	5.82
5	18	16.59	1.41	17.32	0.68
6	18	18.00	0.00	18.39	-0.39
7	19	19.80	-0.80	22.18	-3.18
8	20	19.80	0.20	22.18	-2.18
9	15	16.37	-1.37	16.51	-1.51
10	14	12.70	1.30	12.86	1.14
11	14	12.70	1.30	12.86	1.14
12	13	13.44	-0.44	12.44	0.56
13	11	13.44	-1.70	12.86	-1.86
14	12	13.10	-1.10	14.00	-2.00
15	8	7.34	0.66	8.00	0.00
16	7	7.34	-0.34	8.00	-1.00
17	8	8.00	0.00	8.00	0.00
18	8	8.00	0.00	8.00	0.00
19	9	11.81	-2.81	9.37	-0.37
20	15	13.13	1.87	13.15	1.85
21	15	17.06	-2.06	15.00	0.00
			SS = 119.176		SS = 91.677

Figure 17. Predicted Values and Residuals Using Forward Rank Regression and Backward Rank Regression.

References

Brownlee, K. A. (1965). Statistical Theory and Methodology in Science and Engineering, 2nd ed. John Wiley and Sons, Inc., New York.

Conover, W. J. (1980). Practical Nonparametric Statistics, 2nd ed. John Wiley and Sons, Inc., New York.

Draper, N. R. and Smith, H. (1966). Applied Regression Analysis. John Wiley and Sons, Inc., New York.

Iman, Ronald L. and Conover, W. J. (1979). The Use of the Rank Transform in Regression. Technometrics, Vol. 21, No. 4, pp. 499-509.

Iman, Ronald L. and Conover, W. J. (1980). Risk Methodology for Geologic Disposal of Radioactive Waste: A Distribution Free Approach to Inducing Rank Correlation Among Input Variables for Simulation Studies, NUREG/CR-1262.

Iman, Ronald L., Davenport, James M., Frost, Elizabeth L. and Shortencarier, Michael J. (1980). Stepwise Regression with PRESS and Rank Regression (Program User's Guide). Technical Report, SAND79-1472, Sandia Laboratories, Albuquerque.

Daniel, Cuthbert, and Wood, Fred S. (1971). Fitting Equations to Data. John Wiley & Sons, Inc., New York.

EXAMPLE OF SETTING UP AND EXECUTING THE LATIN HYPERCUBE SAMPLING PROGRAM ALONG WITH OUTPUT FROM A TRANSPORT MODEL

This part of the course demonstrates the techniques of the first two sections on the NWFT/DVM model. The NWFT/DVM model uses 17 input variables, 4 of which are correlated.

Page 104 shows the parameter cards used to generate the Latin hypercube sample as described in SAND79-1473.

Page 105 gives the user specified subroutine as required on cards 14 to 21 on page 104.

Pages 106-108 give the actual LHS for 17 variables and a sample of size 35.

Pages 109-111 give the ranks from 1 to 35 for each of the 17 input variables on pages 106-108.

Page 112 contains the output from running the 35 input vectors through NWFT/DVM. The output is in the form of total integrated discharge for 7 isotopes over 10^4 years on a per vector basis.

PARAMETER CARDS USED TO GENERATE A LATIN HYPERCUBE SAMPLE AS DESCRIBED IN THE PROGRAM USER's GUIDE

Card No.

```
6245763762631657                    LHS SCENARIOS  NWFT DVM
TITLE-LHS NWFT DVM FOR NRC SHORT COURSE
   17    35     1     1     0
LOGNORMAL                KD FOR CM(AM)
         .01       1.E5
LOGNORMAL                KD FOR PU
         .01       1.E4
LOGNORMAL                KD FOR U
         .01       1.E4
LOGNORMAL                KD FOR TH
         .01       1.E4
LOGNORMAL                KD FOR NP
         .01        50.
USER-INPUT               SOL LIMIT FOR PU(LOG 10)
        -7.1         2.
USER-INPUT               SOL LIMIT FOR U(LOG 10)
        -5.7         1.
USER-INPUT               SOL LIMIT FOR TH(LOG 10)
        -7.1         .6
USER-INPUT               SOL LIMIT FOR NP(LOG 10)
       -14.4         3.
UNIFORM                  DISPERSIVITY
    1
               50.        500.
LOGUNIFORM               LEACH TIME
    1
             1.E03        1.EC7
LOGNORMAL                K(UPPER AQUIFER)
         .01        50.
NORMAL                   POROSITY(UPPER AQUIFER)
         .05        .30
LOGUNIFORM               K OF THE FEATURE(S)
    1
    1          .05          25.
NORMAL                   POROSITY OF THE FEATURE(S)
         .05        .30
UNIFORM                  TIME OF ONSET OF MIGRATION
    1
             100.         1.E4
UNIFORM                  NUMBER OF ROOMS
    1
              1.0         1100.
    4
   12    13    14    15
    1.0   0.7   1.0   0.0   0.0   1.0   0.0   0.0
    0.7   1.0
OUTPUT,DATA,PLOT,CORR
```

EXAMPLE OF A USER SUPPLIED SUBROUTINE FOR USE WITH THE LATIN HYPERCUBE SAMPLING PROGRAM TO GENERATE A SAMPLE FROM A DISTRIBUTION NOT INCLUDED IN THE PROGRAM - THE IMPLEMENTATION OF THIS SUBROUTINE IS EXPLAINED IN THE PROGRAM USER'S GUIDE

```
      SUBRCUTINE USRDIST(I,N,INSET,IRS,L1,L2,L3)
      COMMON L(10000)
      COMMCN/A/X(100000)
      COMMON/B/XX(100000)
      LEVEL 2,X,XX
      LOC(I,J)=(J-1)*N+I
      ALPINC=0.95/N
      READ(8,110)A,B
  110 FORMAT(2G10.4)
C   A IS THE MEAN ON A LOG10 SCALE FOR NORMAL DISTRIBUTION
C   B IS THE ST DEV ON A LOG10 SCALE FOR A NORMAL DISTRIBUTION
C   FOR THIS DISTRIBUTION A WILL BE TREATED AS THE 0.025 QUANTILE
C   AND B WILL BE TREATED AS THE 0.975 QUANTILE
      DELTA=.025
      DO 10 K=1,N
      R=ALPINC*UDGEN(0.)+DELTA
      X(LOC(K,I))=10.**(FINVNOR(R)*B+A)
      DELTA=DELTA+ALPINC
   10 CONTINUE
      IF(INSET.NE.1)GO TO 20
      PRINT 120,I,A,B
      PRINT 130,L1,L2,L3
   20 CONTINUE
  120 FORMAT(*C*,11X,I3,12X,*USER SUPPLIED MEAN=*,1PG10.3,
     *       *ST DEV=*,1PG10.3)
  130 FORMAT(1H+,75X,3A10)
      RETURN
      END
```

TITLE-LHS NWFT DVM FOR NRC SHORT COURSE

INPUT VECTORS

RUN NO.	X(1)	X(2)	X(3)	X(4)	X(5)	X(6)	X(7)	X(8)	X(9)	X(10)	X(11)	X(12)
1	943.	22.1	9.06	35.3	1.22	1.900E-11	5.084E-08	6.738E-08	3.049E-13	314.	1.152E+03	7.913E-02
2	152.	55.5	4.54	.770	9.440E-02	6.981E-06	7.714E-06	3.695E-08	8.463E-15	280.	4.481E+03	.205
3	316.	533.	21.0	5.31	2.83	2.548E-08	9.521E-06	2.914E-08	2.782E-18	191.	2.076E+04	4.51
4	9.12	3.24	1.15	1.17	1.30	3.469E-10	2.216E-07	5.972E-08	7.788E-10	263.	5.538E+04	.355
5	3.783E+03	1.36	3.92	29.6	1.64	8.716E-05	6.047E-06	5.459E-07	5.060E-11	339.	2.077E+06	8.83
6	14.5	20.6	17.5	43.5	.484	8.976E-09	6.158E-07	2.090E-07	1.361E-18	282.	2.276E+03	12.6
7	5.51	36.9	87.2	.235	2.37	4.347E-08	4.851E-05	5.642E-07	8.737E-16	352.	1.739E+03	1.05
8	1.34	6.82	65.5	372.	.170	2.247E-09	8.819E-08	1.316E-07	6.134E-16	145.	1.489E+06	.580
9	20.6	85.0	250.	3.01	1.53	3.456E-11	4.611E-06	1.460E-07	1.465E-10	160.	4.639E+05	.831
10	2.128E+04	.782	5.89	188.	4.747E-02	1.359E-09	1.626E-06	1.969E-07	3.465E-19	388.	2.999E+04	1.51
11	29.0	1.032E+03	.611	22.8	.665	1.771E-08	2.384E-06	1.049E-07	1.343E-16	62.6	6.585E+05	.116
12	1.328E+03	180.	174.	18.6	.362	1.011E-09	4.229E-07	5.657E-08	2.419E-13	398.	1.632E+06	1.14
13	47.9	11.6	9.10	1.001E+04	2.01	8.471E-08	3.127E-06	1.918E-08	7.219E-12	493.	2.356E+05	2.66
14	195.	5.09	6.85	.318	.791	1.132E-06	1.016E-05	3.474E-07	3.806E-17	248.	3.969E+06	.316
15	18.1	4.60	1.653E+03	55.9	.906	2.170E-06	3.583E-07	4.140E-07	2.427E-14	63.9	8.630E+03	.145
16	98.2	338.	2.60	87.2	.260	5.159E-08	1.264E-04	3.813E-08	2.644E-12	361.	7.106E+03	.187
17	8.79	16.6	1.65	4.43	.394	8.706E-07	2.785E-06	7.969E-09	1.254E-12	92.2	2.824E+05	1.82
18	2.55	2.02	142.	330.	.516	3.209E-09	1.705E-05	2.502E-08	1.862E-16	202.	7.084E+06	.709
19	3.26	10.4	13.4	1.61	1.88	1.171E-08	8.056E-05	1.414E-08	4.072E-16	451.	3.773E+05	1.68
20	71.4	145.	18.0	1.23	6.21	1.249E-07	2.655E-07	7.142E-08	2.899E-17	323.	1.162E+06	1.28
21	1.87	51.1	29.4	10.8	.594	4.222E-04	4.291E-08	4.982E-08	1.785E-11	445.	3.138E+04	.248
22	52.2	.488	43.1	35.9	9.01	1.129E-05	2.652E-05	3.013E-07	4.215E-14	429.	1.517E+05	3.619E-02
23	85.7	6.56	2.02	2.45	.288	4.321E-06	8.858E-07	2.060E-08	1.530E-20	417.	1.191E+05	.278
24	.418	.867	75.5	12.6	.135	2.086E-05	1.111E-06	9.645E-09	2.264E-15	300.	1.485E+04	2.27
25	400.	.134	11.5	3.48	4.08	1.948E-07	1.234E-06	4.479E-08	4.125E-15	78.8	1.251E+04	5.75
26	.636	1.74	.900	2.20	.915	2.495E-10	1.858E-07	1.825E-07	1.768E-15	475.	3.899E+06	.409
27	138.	8.09	.375	13.1	1.02	1.376E-06	1.151E-07	1.648E-07	8.978E-17	473.	3.430E+03	.469
28	4.40	2.73	4.566E-02	17.0	30.7	5.148E-09	7.221E-07	8.201E-08	6.490E-18	113.	7.866E+05	.653
29	6.33	67.3	34.8	135.	3.45	2.893E-05	2.141E-06	1.205E-07	2.068E-19	242.	4.998E+04	.914
30	24.7	.371	341.	.617	.215	6.690E-10	4.048E-06	3.177E-08	1.630E-17	378.	1.123E+05	8.114E-02

ACTUAL LATIN HYPERCUBE SAMPLE GENERATED

31	11.1	28.5	1.26	7.53	7.578E-02	4.083E-07	1.483E-05	9.187E-07	5.662E-15	224.	8.492E+06	.942
32	586.	30.2	5.04	6.43	.685	0.679E-07	1.471E-06	1.509E-08	7.960E-14	132.	5.073E+06	.528
33	250.	4.02	37.1	1.872E-02	.113	3.042E-07	4.911E-07	8.457E-08	1.030E-13	212.	8.608E+04	2.09
34	.160	13.1	.206	8.43	.232	1.632E-07	5.397E-06	2.489E-07	8.395E-13	172.	1.587E+03	3.24
35	35.2	.277	2.83	68.8	.442	2.825E-08	2.023E-05	9.739E-08	1.888E-14	116.	6.135E+03	.392

TITLE-LHS NWFT DVM FOR NRC SHORT COURSE

INPUT VECTORS

RUN NO.	X(13)	X(14)	X(15)	X(16)	X(17)
1	.123	3.03	.168	1.511E+03	132.
2	.147	.138	.110	6.379E+03	1.090E+03
3	.164	18.3	.201	813.	1.038E+03
4	.169	9.332E-02	.162	4.576E+03	719.
5	.215	.848	.165	5.687E+03	464.
6	.269	.105	.118	5.242E+03	524.
7	.203	6.147E-02	.147	3.228E+03	655.
8	.190	.566	.153	7.517E+03	735.
9	.181	.500	.127	9.259E+03	54.5
10	.141	7.876E-02	.123	8.770E+03	320.
11	.168	1.57	.230	8.499E+03	813.
12	.212	6.42	.217	9.947E+03	912.
13	.231	.176	.179	1.736E+03	601.
14	.117	.154	.158	7.912E+03	496.
15	.105	.225	.142	1.845E+03	998.
16	.161	2.20	.202	4.128E+03	364.
17	.196	5.038E-02	.137	3.761E+03	189.
18	.127	.355	.211	350.	413.
19	.155	3.92	.174	8.917E+03	560.
20	.192	12.3	.239	6.814E+03	26.3
21	7.728E-02	1.84	.195	6.991E+03	106.
22	.151	.877	.182	7.235E+03	573.
23	.187	.272	.173	2.512E+03	91.9
24	.206	2.82	.193	9.585E+03	926.
25	.242	4.74	.183	5.181E+03	198.
26	.175	1.30	.136	4.699E+03	958.
27	.132	8.11	.223	6.258E+03	868.
28	.157	.320	.161	5.774E+03	818.
29	.227	1.17	.152	3.117E+03	388.
30	.144	5.46	.187	4.055E+03	231.
31	.194	16.8	.190	627.	293.
32	.178	.642	7.814E-02	2.101E+03	678.
33	.220	10.4	.234	1.119E+03	1.019E+03
34	.200	8.71	.208	2.754E+03	256.
35	.172	24.6	.255	8.193E+03	772.

TITLE-LHS NWFT DVM FOR VRC SHOFT COURSE

RANKS OF INPUT VECTORS

RUN NO.	X(1)	X(2)	X(3)	X(4)	X(5)	X(6)	X(7)	X(8)	X(9)	X(10)	X(11)	X(12)
1	32.	23.	18.	25.	23.	1.	2.	16.	27.	21.	1.	2.
2	26.	29.	13.	5.	3.	30.	26.	10.	20.	18.	6.	7.
3	29.	34.	23.	14.	30.	14.	27.	8.	5.	11.	12.	32.
4	12.	11.	6.	6.	24.	4.	6.	15.	35.	17.	16.	11.
5	34.	7.	12.	24.	26.	34.	25.	33.	33.	23.	30.	34.
6	14.	22.	21.	27.	14.	11.	11.	28.	4.	19.	4.	35.
7	9.	26.	30.	2.	29.	16.	33.	34.	15.	24.	3.	22.
8	4.	16.	28.	34.	6.	9.	3.	23.	14.	8.	28.	16.
9	16.	30.	33.	11.	25.	2.	23.	24.	34.	9.	24.	20.
10	35.	5.	15.	32.	1.	7.	17.	27.	3.	27.	13.	25.
11	18.	35.	4.	23.	17.	13.	19.	21.	11.	1.	25.	4.
12	33.	32.	32.	22.	11.	6.	9.	14.	26.	26.	29.	23.
13	20.	19.	17.	35.	28.	18.	21.	5.	31.	35.	21.	30.
14	27.	14.	16.	3.	19.	26.	28.	31.	9.	16.	31.	10.
15	15.	13.	35.	28.	20.	28.	8.	32.	22.	2.	9.	5.
16	24.	33.	10.	30.	9.	17.	35.	11.	30.	25.	8.	6.
17	11.	21.	8.	13.	12.	25.	20.	1.	29.	4.	22.	27.
18	6.	9.	31.	33.	15.	9.	30.	7.	12.	12.	34.	18.
19	7.	19.	20.	8.	27.	12.	34.	3.	13.	32.	23.	26.
20	22.	31.	22.	7.	33.	19.	7.	17.	8.	22.	27.	24.
21	5.	27.	24.	18.	16.	35.	1.	13.	32.	31.	14.	8.
22	21.	4.	27.	26.	34.	31.	32.	30.	23.	30.	20.	1.
23	23.	15.	9.	10.	10.	29.	13.	6.	1.	29.	19.	9.
24	2.	6.	29.	19.	5.	32.	14.	2.	17.	20.	11.	29.
25	30.	1.	19.	12.	32.	21.	15.	12.	18.	3.	10.	33.
26	3.	8.	5.	9.	21.	3.	5.	26.	16.	34.	32.	13.
27	25.	17.	3.	20.	22.	27.	4.	25.	10.	33.	5.	14.
28	8.	10.	1.	21.	35.	10.	12.	18.	6.	5.	26.	17.
29	10.	29.	25.	31.	31.	33.	18.	22.	2.	15.	15.	19.
30	17.	3.	34.	4.	7.	5.	22.	9.	7.	26.	18.	3.

CORRESPONDING RANKS OF THE VARIABLES IN THE LATIN HYPERCUBE SAMPLE

31	13.	24.	7.	16.	2.	23.	29.	35.	19.	14.	35.	21.
32	31.	25.	14.	15.	18.	24.	16.	4.	24.	7.	33.	15.
33	28.	12.	26.	1.	4.	22.	10.	19.	25.	13.	17.	28.
34	1.	20.	2.	17.	8.	20.	24.	29.	28.	10.	2.	31.
35	19.	2.	11.	29.	13.	15.	31.	20.	21.	6.	7.	12.

TITLE-LHS NWFT DVM FOR NRC SHORT COURSE

RANKS OF INPUT VECTORS

RUN NO.	X(13)	X(14)	X(15)	X(16)	X(17)
1	4.	24.	16.	5.	5.
2	9.	6.	2.	23.	35.
3	14.	34.	26.	3.	34.
4	16.	4.	14.	16.	23.
5	30.	15.	15.	20.	15.
6	35.	5.	3.	19.	17.
7	27.	2.	9.	12.	21.
8	23.	14.	11.	27.	24.
9	20.	13.	5.	33.	2.
10	7.	3.	4.	31.	11.
11	15.	20.	32.	30.	26.
12	29.	29.	30.	35.	29.
13	33.	8.	19.	6.	20.
14	3.	7.	12.	28.	16.
15	2.	9.	8.	7.	32.
16	13.	22.	27.	15.	12.
17	25.	1.	7.	13.	6.
18	5.	12.	29.	1.	14.
19	11.	25.	18.	32.	18.
20	24.	32.	34.	24.	1.
21	1.	21.	25.	25.	4.
22	10.	17.	20.	26.	19.
23	22.	10.	17.	9.	3.
24	28.	23.	24.	34.	30.
25	34.	26.	21.	18.	7.
26	18.	19.	6.	17.	31.
27	6.	29.	31.	22.	28.
28	12.	11.	13.	21.	27.
29	32.	18.	10.	11.	13.
30	8.	27.	22.	14.	8.
31	21.	33.	23.	2.	10.
32	19.	15.	1.	8.	22.
33	31.	31.	33.	4.	33.
34	26.	30.	28.	10.	9.
35	17.	35.	35.	29.	25.

SAMPLE OUTPUT GENERATED FROM A TRANSPORT MODEL USING THE PREVIOUS LATIN HYPERCUBE SAMPLE AS INPUT

Total Integrated Discharge to 10^4 Years

Run No.	^{237}Np	^{233}U	^{229}Th	^{246}Cm	^{242}Pu	^{238}U	^{238}Pu
1	0.	0.	0.	0.	0.	0.	0.
2	4.809E-05	3.997E-09	0.	0.	0.	0.	0.
3	5.807E-02	2.110E-04	8.775E-03	0.	0.	3.097E-05	0.
4	4.066E-07	3.740E-08	1.311E-06	0.	0.	2.754E-10	0.
5	8.083E-06	1.329E-04	2.863E-06	0.	2.385E-04	1.268E-05	0.
6	5.576E-04	1.214E-06	1.101E-06	2.403E-02	9.049E-05	1.187E-07	0.
7	7.907E-05	2.207E-08	8.621E-04	6.525E-05	1.710E-08	0.	0.
8	9.516E-07	1.721E-11	0.	0.	0.	0.	0.
9	0.	0.	0.	0.	0.	0.	0.
10	3.740E-06	6.712E-13	0.	0.	5.496E-08	0.	0.
11	0.	0.	0.	0.	0.	0.	0.
12	0.	0.	0.	0.	0.	0.	0.
13	3.311E-05	2.807E-06	3.090E-09	0.	4.503E-06	7.751E-07	0.
14	0.	0.	1.054E-07	0.	0.	0.	0.
15	0.	0.	0.	0.	0.	0.	0.
16	8.823E-04	5.075E-07	0.	0.	0.	0.	0.
17	6.064E-06	5.218E-06	5.379E-07	4.088E-10	0.	6.111E-07	0.
18	4.677E-06	7.400E-10	5.223E-11	7.345E-06	1.052E-06	0.	0.
19	0.	0.	0.	0.	0.	0.	0.
20	1.139E-06	2.458E-06	6.329E-05	6.568E-05	6.829E-06	2.269E-07	0.
21	0.	0.	0.	0.	0.	0.	0.
22	0.	0.	0.	0.	0.	0.	0.
23	3.175E-05	5.000E-08	2.164E-09	0.	0.	2.289E-10	0.
24	1.133E-05	0.	0.	2.354E-03	3.387E-07	0.	0.
25	4.759E-03	3.276E-05	4.575E-03	0.	1.372E-02	2.614E-06	0.
26	1.227E-06	7.218E-07	4.092E-08	1.737E-04	3.003E-07	8.495E-08	0.
27	4.045E-05	1.882E-05	5.402E-08	0.	0.	9.383E-07	0.
28	0.	1.138E-05	1.549E-08	0.	0.	1.077E-06	0.
29	2.666E-05	1.888E-08	0.	0.	0.	0.	0.
30	4.007E-05	1.443E-09	0.	0.	1.936E-07	0.	0.
31	1.215E-04	2.401E-04	8.373E-06	0.	0.	3.816E-05	0.
32	2.404E-06	0.	0.	0.	0.	0.	0.
33	1.526E-02	1.349E-06	2.725E-02	0.	8.097E-03	0.	0.
34	8.162E-02	5.781E-03	3.261E-03	12.2	1.471E-03	6.204E-04	0.
35	0.	0.	0.	0.	6.242E-06	0.	0.

AN EXAMPLE OF SENSITIVITY ANALYSIS RESULTS BASED ON THE PARTIAL RANK CORRELATION COEFFICIENT

Page 114 contains a summary of the partial rank correlation coefficients as given in Equation (19) on page 98 on Part Two of this course. Listed down the left hand side of the table are the 17 input variables from page 104. The heading across the top of the columns identifies 6 output variables, 3 of which are listed on page 112. The numerical entries in in the body of the table can be used to identify the input variables which are dominate in influencing the output. For example the entry of 70 for KD U and U233 means that the absolute value of the partial rank correlation coefficient was at least .70 at sometime during the 10^4 year period.

While page 114 shows the PRCCs based on the 35 input vectors on pages 106-108, pages 115 and 116 show results for two additional sets of 35 input vectors. The results are similar on pages 114, 115, and 116.

Page 117 shows the summary for all 105 input vectors pooled together but the individual numerical entries show a PRCC of at least .50 rather than .70 as on pages 114-116. On page 118 the 105 vector results are repeated for PRCCs of at least .70.

Pages 119-132 provide PRCC plots for each combination of input variable and output variable that have numerical table entries on page 117.

WHERE TABLE ENTRIES OCCUR THE PRCC BETWEEN THE INPUT VARIABLE (ROW) AND THE OUTPUT VARIABLE (COLUMN) ACHIEVED AT LEAST THE LEVEL .7 OR .8 IN ABSOLUTE VALUE AS INDICATED AT SOME POINT IN TIME OVER THE 10,000 YEAR PERIOD

PRCC LHS-NUFT-DV4 VS TID-S3-CH2 HLW(1.E4 YRS) NRC SHORT COURSE

Input Variables	ISOTOPE	Output Variables CM 245	PU 241	AM 241	NP 237	U 233	TH 229
1.	KD CM(AM)	80		70			
2.	KD PU						
3.	KD U					70	70
4.	KD TH						80
5.	KD NP						
6.	SOL LIM PU						
7.	SOL LIM U						
8.	SOL LIM TH						
9.	SOL LIM NP						
10.	DISPERSIV						
11.	LEACH TIME						
12.	K UP AQ						70
13.	POR UP AQ						
14.	K FEAT						
15.	POR FEAT						
16.	REL TIME	70			80	80	80
17.	NUM RMS						

A TABLE SIMILAR TO THE ONE ON PAGE 114 BUT WITH A NEW LATIN HYPERCUBE SAMPLE FOR DEMONSTRATING CONSISTENCY IN IDENTIFICATION OF IMPORTANT VARIABLES

PRCC LHS-NWFT-DV4 VS TID-S3-CH2 HLW(1.E4 YRS) NRC SHORT COURSE

Input Variables	ISOTOPE	Output Variables: CM 245	PU 241	AM 241	NP 237	U 233	TH 229
1.	KD CM(AM)	80	70	70			
2.	KD PU						
3.	KD U					80	70
4.	KD TH						70
5.	KD NP				70	70	
6.	SOL LIM PU						
7.	SOL LIM U						
8.	SOL LIM TH						
9.	SOL LIM NP						
10.	DISPERSIV						
11.	LEACH TIME				80		
12.	K UP AQ					70	70
13.	POR UP AQ						
14.	K FEAT						
15.	POR FEAT						
16.	REL TIME				80	80	80
17.	NUM RMS						

RESULTS FROM A THIRD LATIN HYPERCUBE
SAMPLE TO COMPARE WITH THE TABLES ON
THE PREVIOUS TWO PAGES

PRCC LHS-NWFT-DVM VS TID-S3-CH2 HLW(1.E4 YRS), NRC SHORT COURSE

Input Variables		Output Variables					
	ISOTOPE	CM 245	PU 241	AM 241	NP 237	U 233	TH 229
1.	KD CM(AM)	70	70	70			
2.	KD PU						
3.	KD U					80	
4.	KD TH						70
5.	KD NP				70		
6.	SOL LIM PU						
7.	SOL LIM U						
8.	SOL LIM TH						
9.	SOL LIM NP						
10.	DISPERSIV						
11.	LEACH TIME						
12.	K UP AQ						70
13.	POR UP AQ						
14.	K FEAT						
15.	POR FEAT						
16.	REL TIME		70	70	80	80	80
17.	NUM RMS						

COMPOSITE RESULTS FROM POOLING ALL COMPUTER RUNS FROM THE THREE PRECEDING LATIN HYPERCUBE SAMPLES WHERE FILTERS WERE LOWERED TO .5 AND .6 DUE TO THE INCREASED SAMPLE SIZE

PRCC LHS-NUFT-DV4 VS TID-S3-CH2 HLW(1.E4 YRS) NRC SHORT COURSE

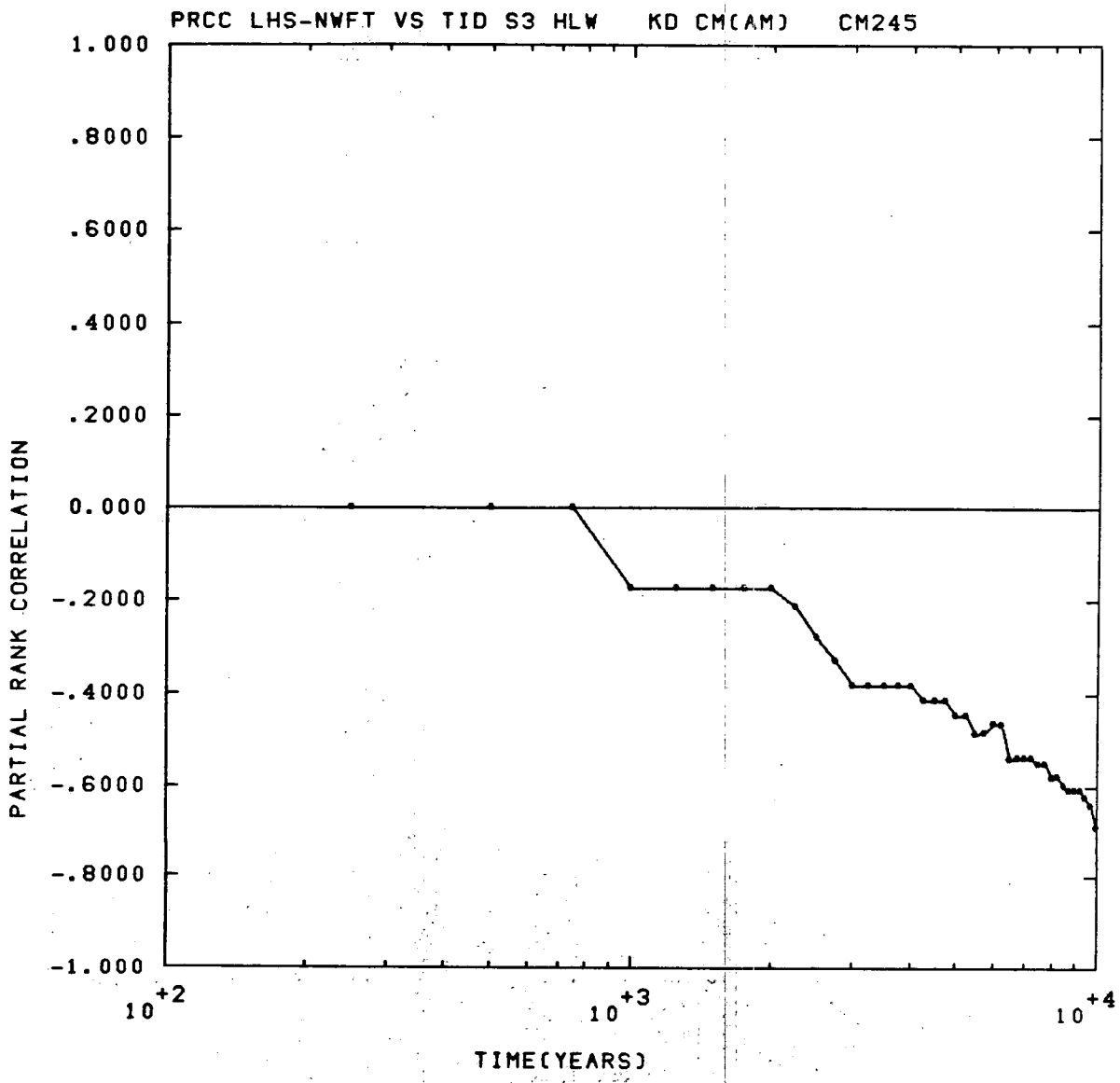
Input Variables		Output Variables					
	ISOTOPE	CM 245	PU 241	AM 241	NP 237	U 233	TH 229
1.	KD CM(AP)	60	50	60			
2.	KD PU						
3.	KD U					60	50
4.	KD TH						60
5.	KD NP				60		
6.	SOL LIM PU						
7.	SOL LIM U						
8.	SOL LIM TH						
9.	SOL LIM NP						
10.	DISPERSIV						
11.	LEACH TIME				60		
12.	K UP AQ				50	50	60
13.	POR UP AQ						
14.	K FEAT						
15.	POR FEAT						
16.	REL TIME				60	60	60
17.	NUM RMS						

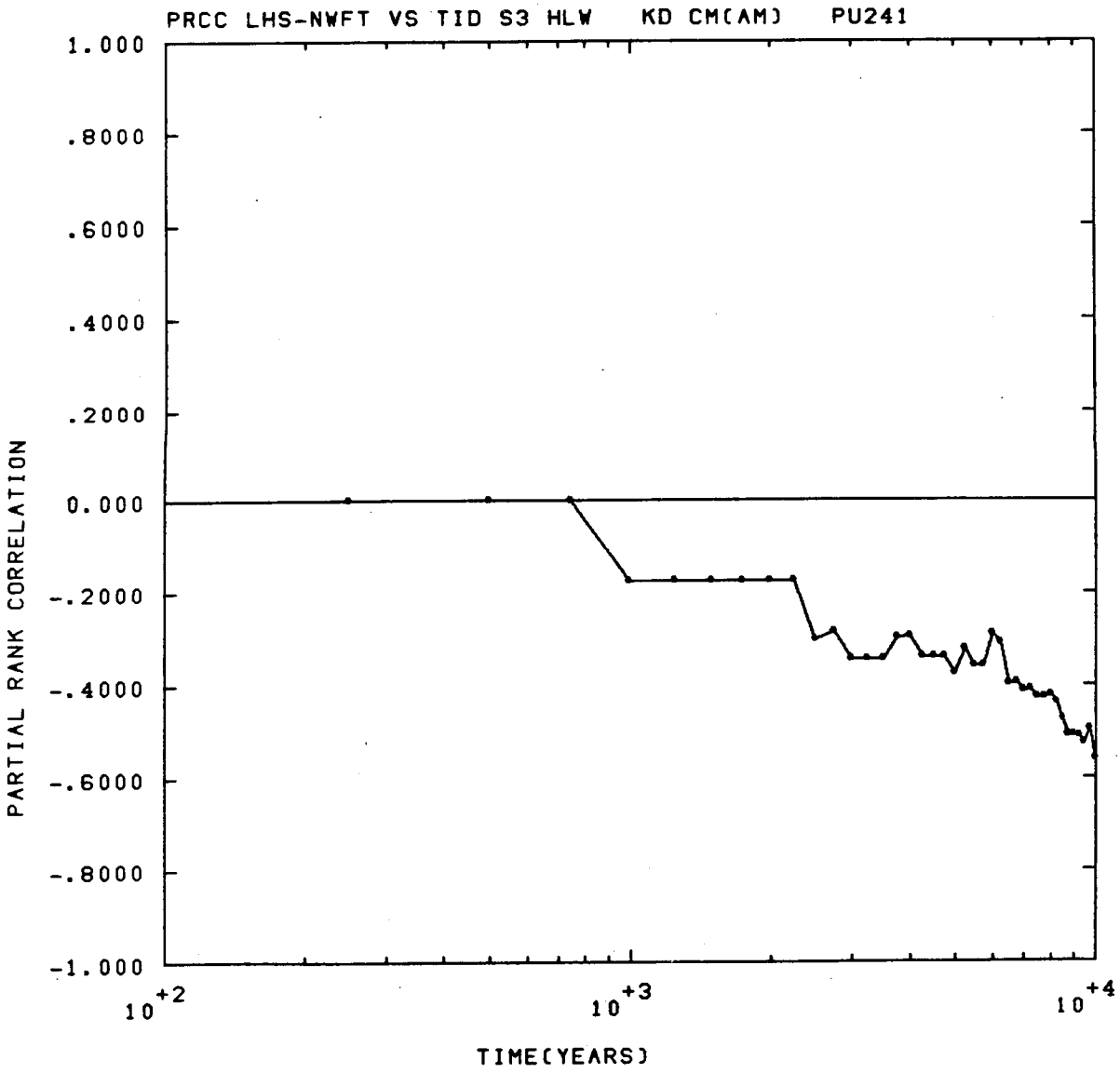
SAME COMPOSITE RESULTS AS ON PREVIOUS PAGE ONLY WITH THE FILTERS INCREASED TO .7 AND .8 FOR PURPOSES ON PINPOINTING DOMINANT VARIABLES

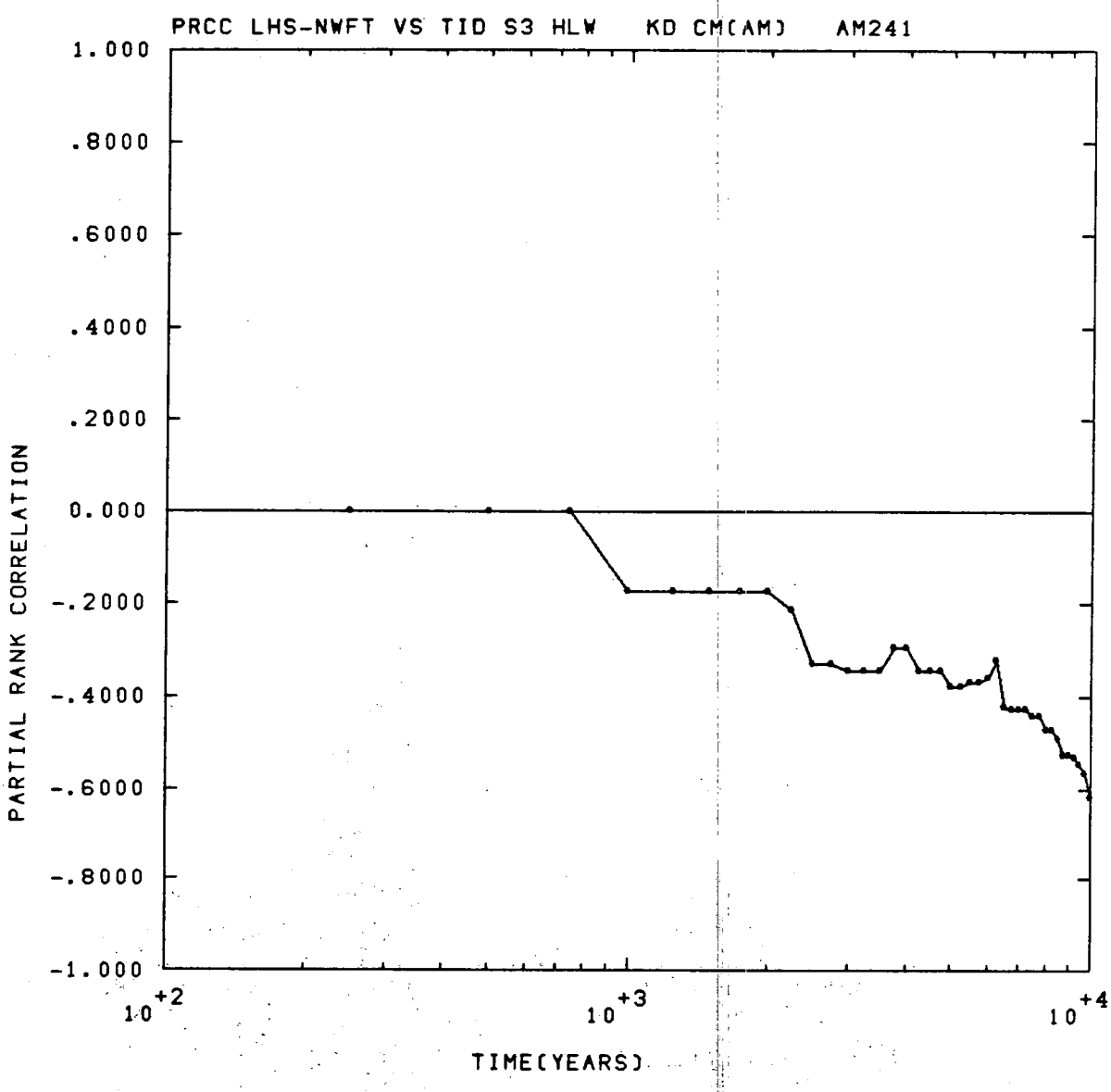
PRCC LHS-NWFT-DVM VS TID-S3-CH2 HLW(1.E4 YRS) NRC SHORT COURSE

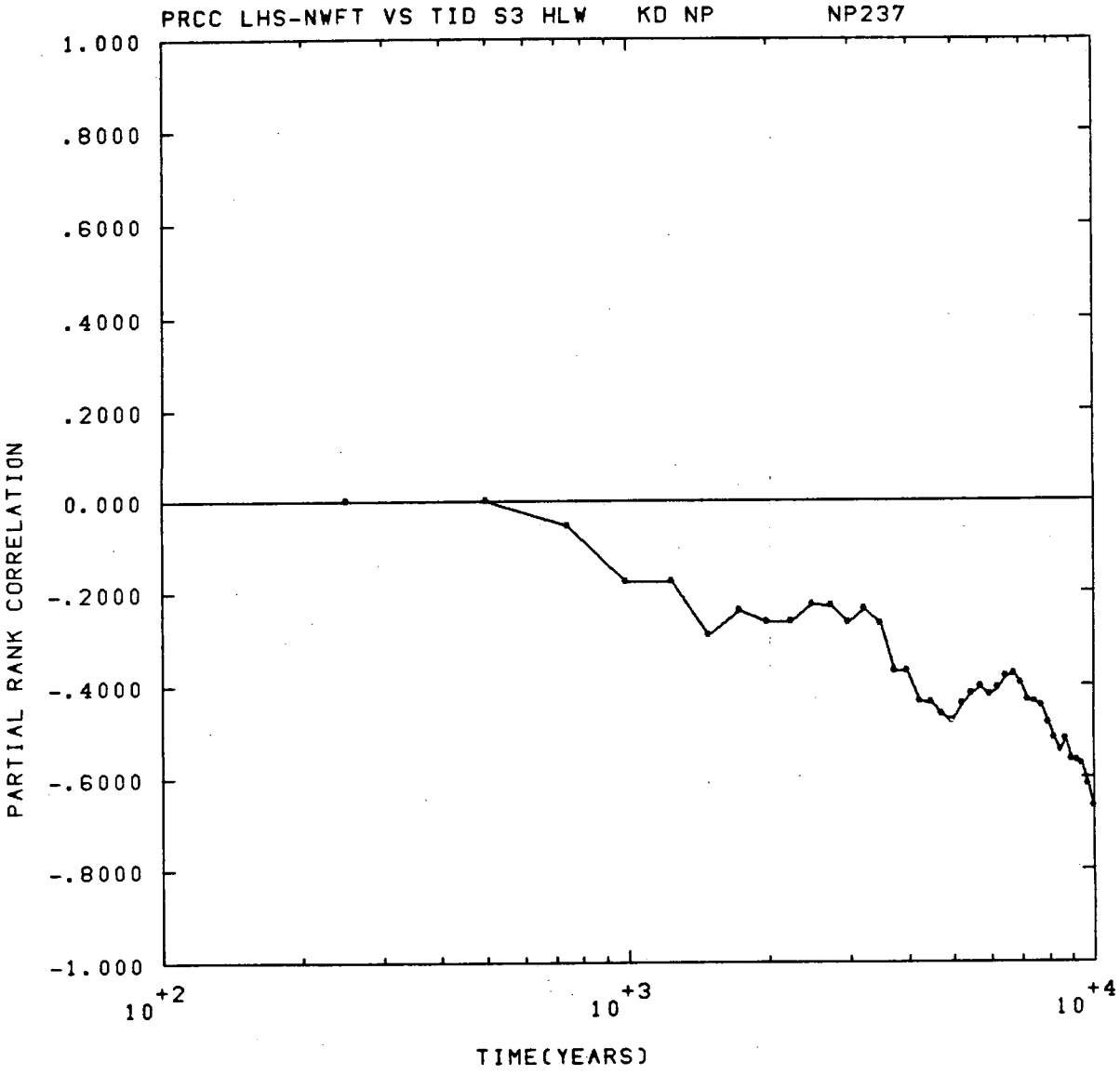
Input Variables	ISOTOPE	Output Variables CM 245	PU 241	AM 241	NP 237	U 233	TH 229
1.	KD CM(AM)						
2.	KD PU						
3.	KD U					70	
4.	KD TH						
5.	KD NP						
6.	SOL LIM PU						
7.	SOL LIM U						
8.	SOL LIM TH						
9.	SOL LIM NP						
10.	DISPERSIV						
11.	LEACH TIME						
12.	K UP AQ						
13.	POR UP AQ						
14.	K FEAT						
15.	POR FEAT						
16.	REL TIME				80	80	80
17.	NUM RMS						

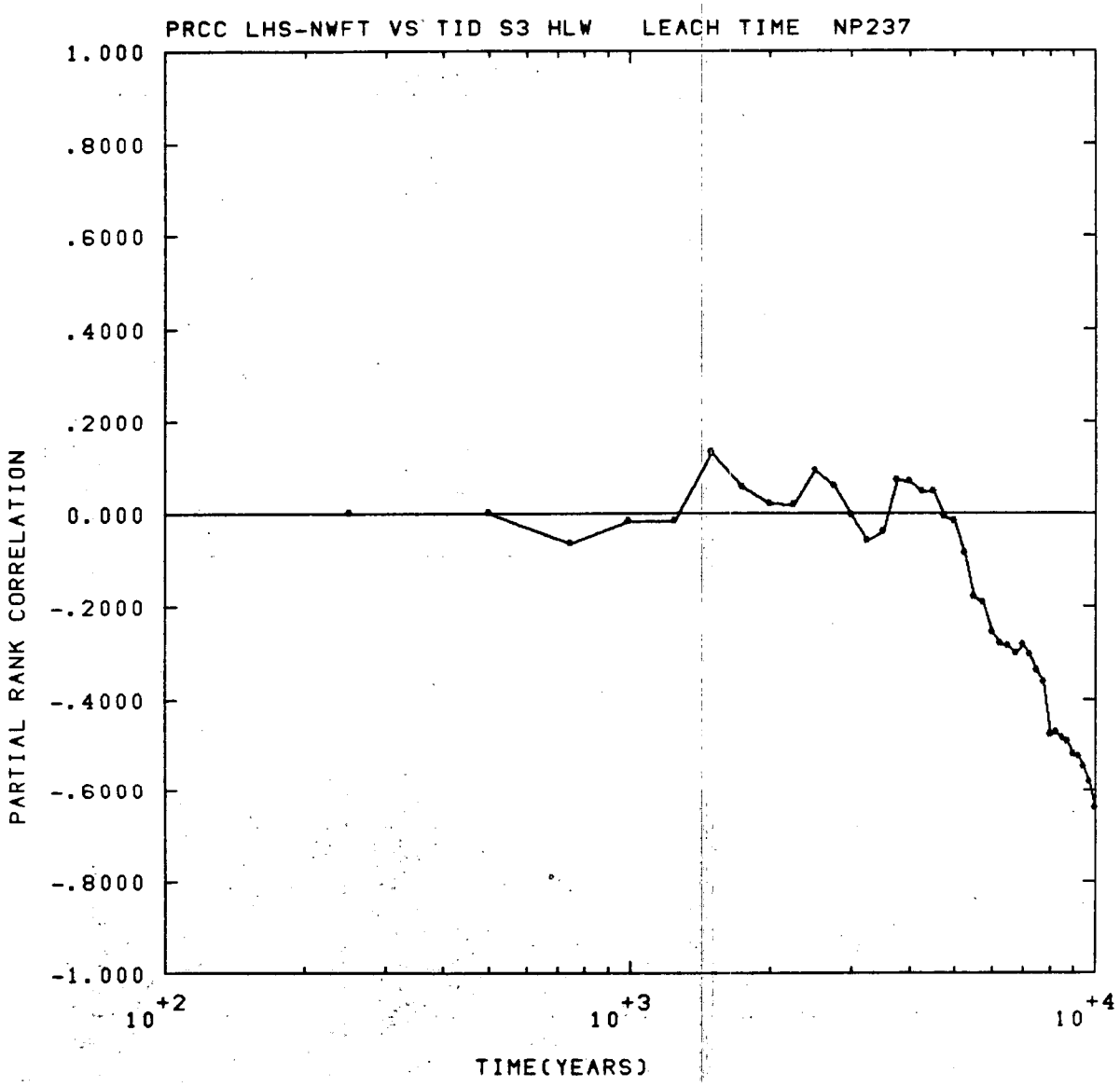
PLOTS OF PRCC's FOR ALL VARIABLES
IDENTIFIED AS IMPORTANT ON PAGE 117

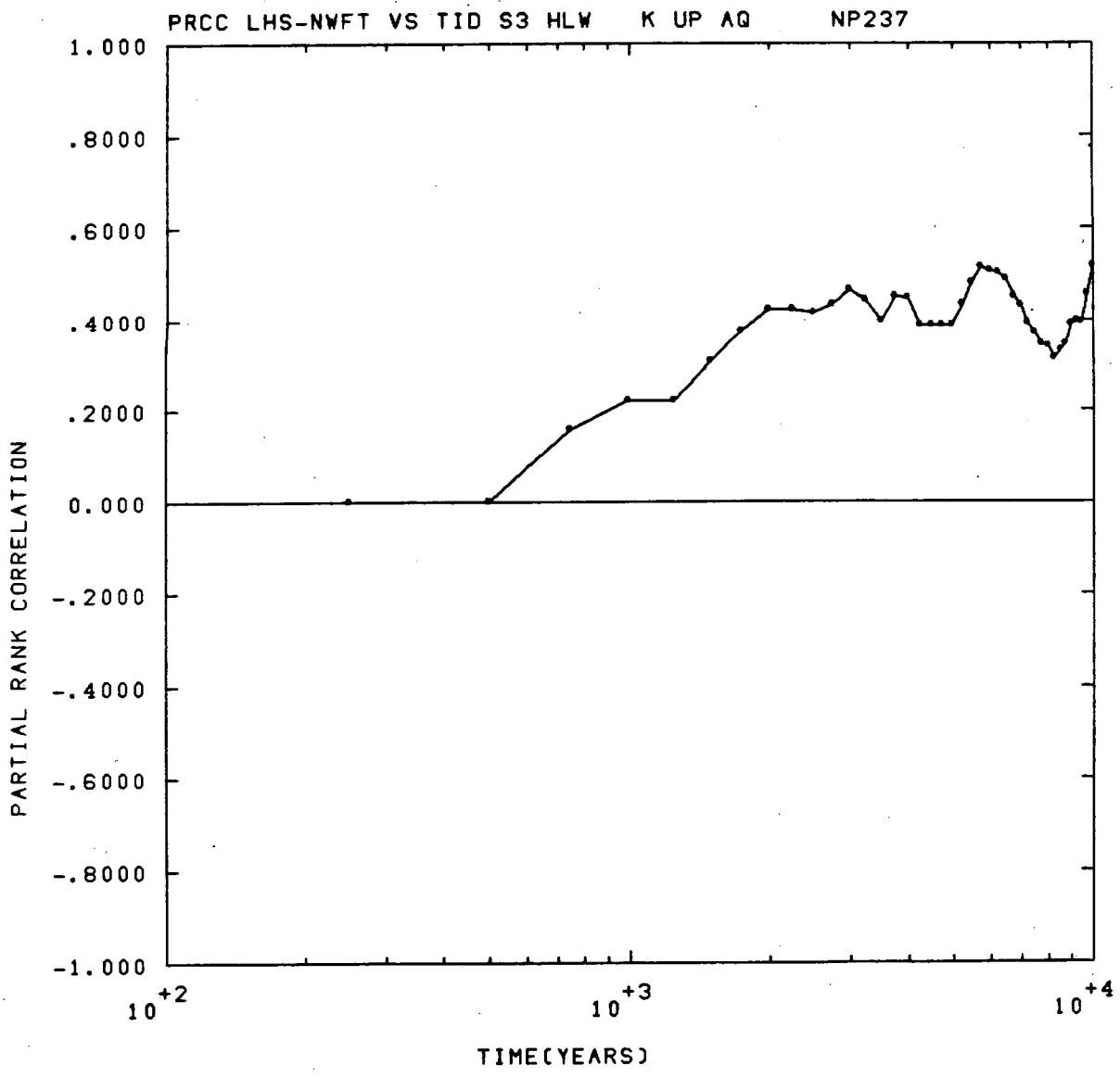


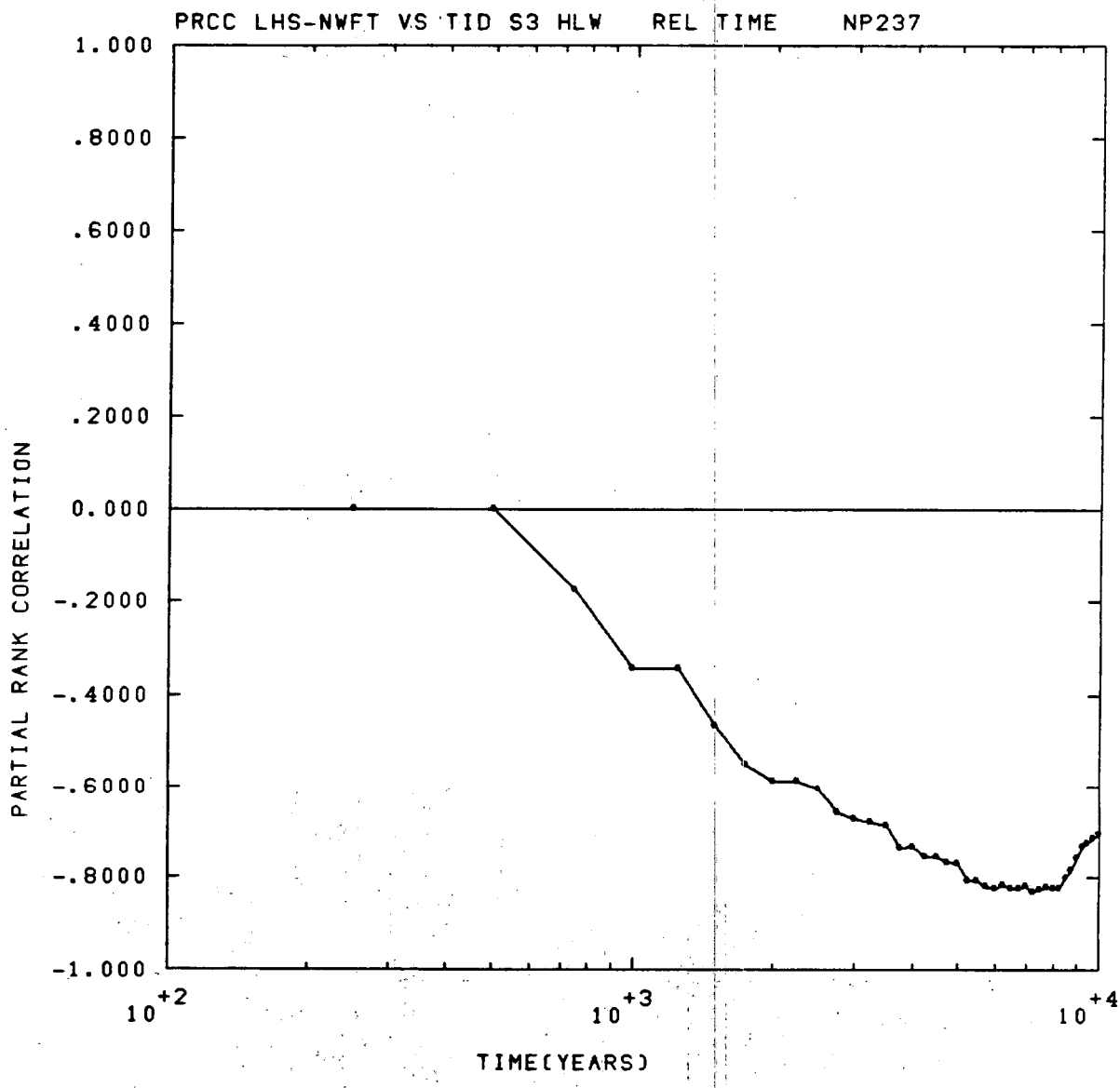

PRCC LHS-NWFT VS TID S3 HLW KD CM(AM) PU241
1.000
.8000
.6000
.4000
.2000
0.000
-.2000
-.4000
-.6000
-.8000
-1.000
PARTIAL RANK CORRELATION
10+2
10+3
10+4
TIME(YEARS)

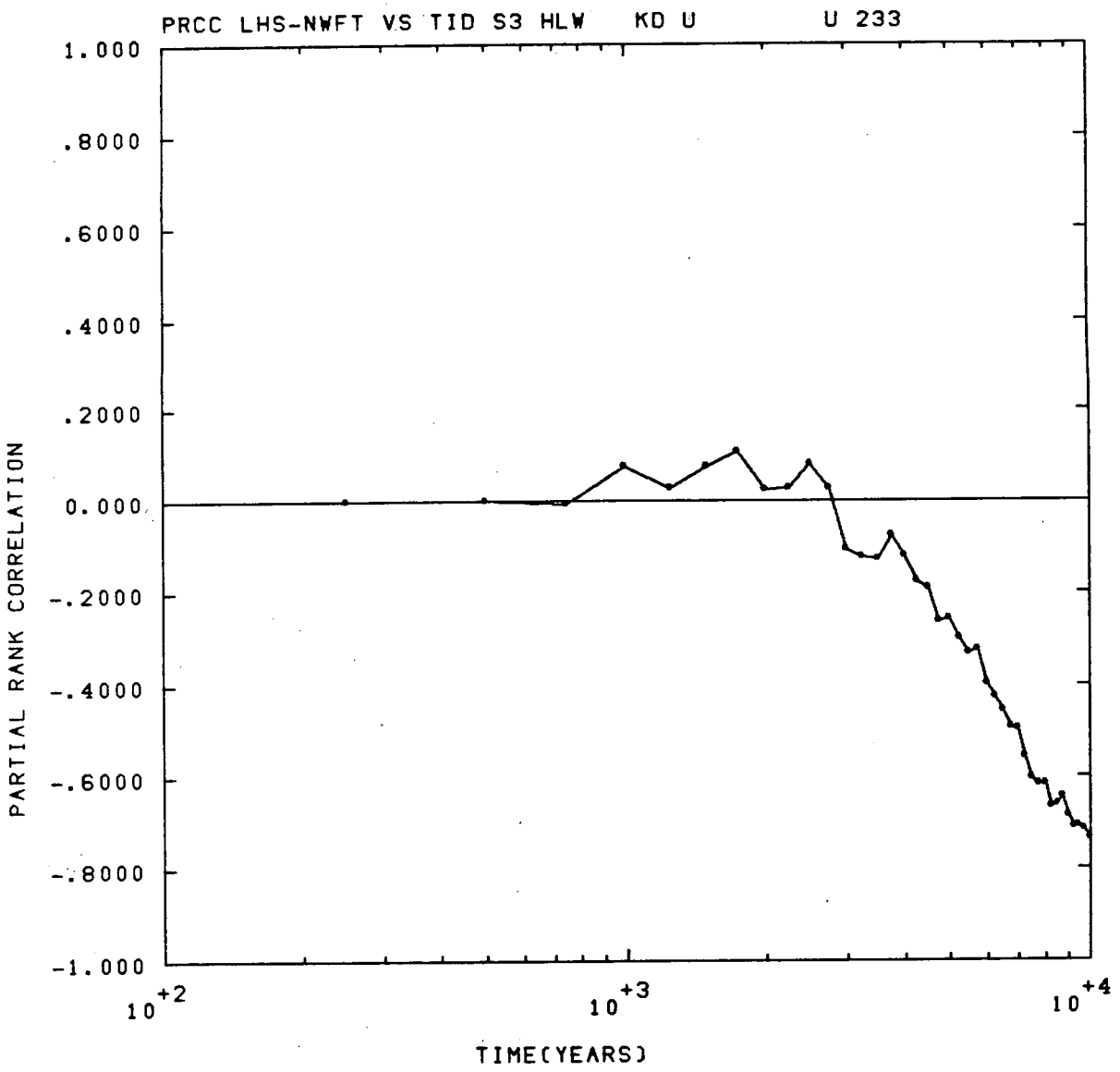

PRCC LHS-NWFT VS TID S3 HLW KD CM(AM) AM241
1.000
.8000
.6000
.4000
.2000
0.000
-.2000
-.4000
-.6000
-.8000
-1.000
PARTIAL RANK CORRELATION
10+2
10+3
10+4
TIME(YEARS)

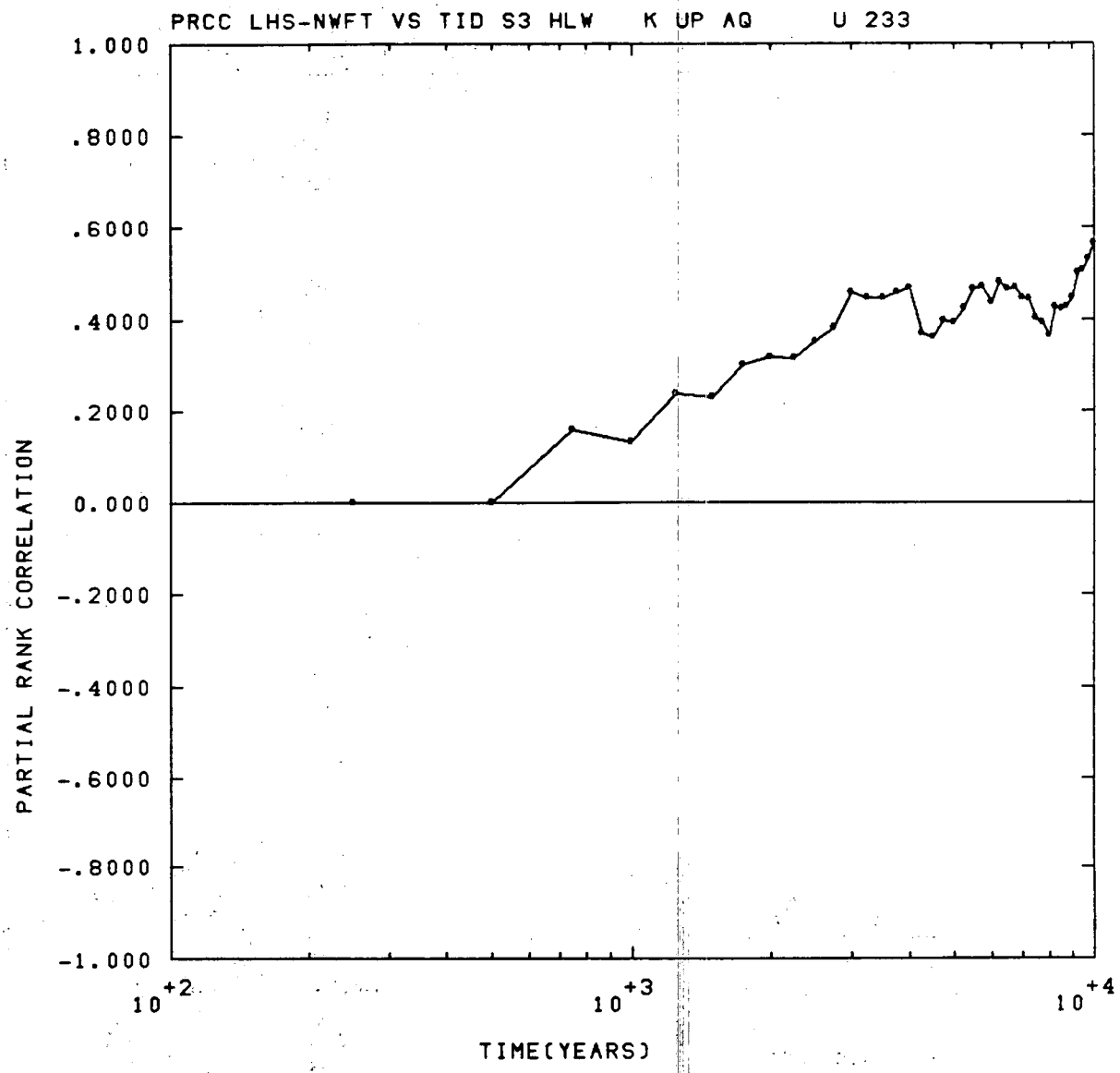

PRCC LHS-NWFT VS TID S3 HLW KD NP NP237
1.000
.8000
.6000
.4000
.2000
0.000
-.2000
-.4000
-.6000
-.8000
-1.000
PARTIAL RANK CORRELATION
10+2
10+3
10+4
TIME(YEARS)

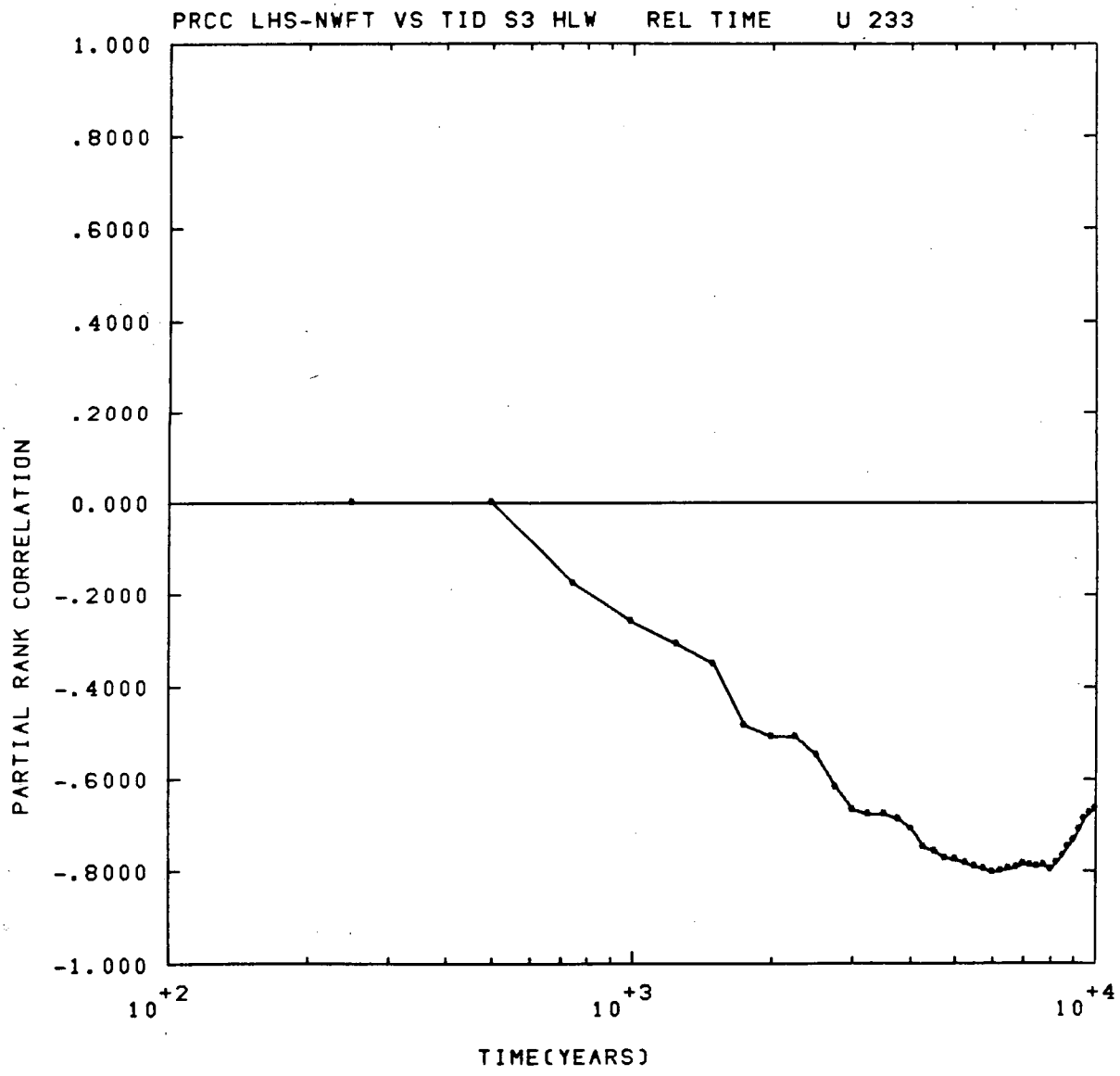

PRCC LHS-NWFT VS TID S3 HLW LEACH TIME NP237
1.000
.8000
.6000
.4000
.2000
0.000
-.2000
-.4000
-.6000
-.8000
-1.000
PARTIAL RANK CORRELATION
10+2
10+3
10+4
TIME(YEARS)

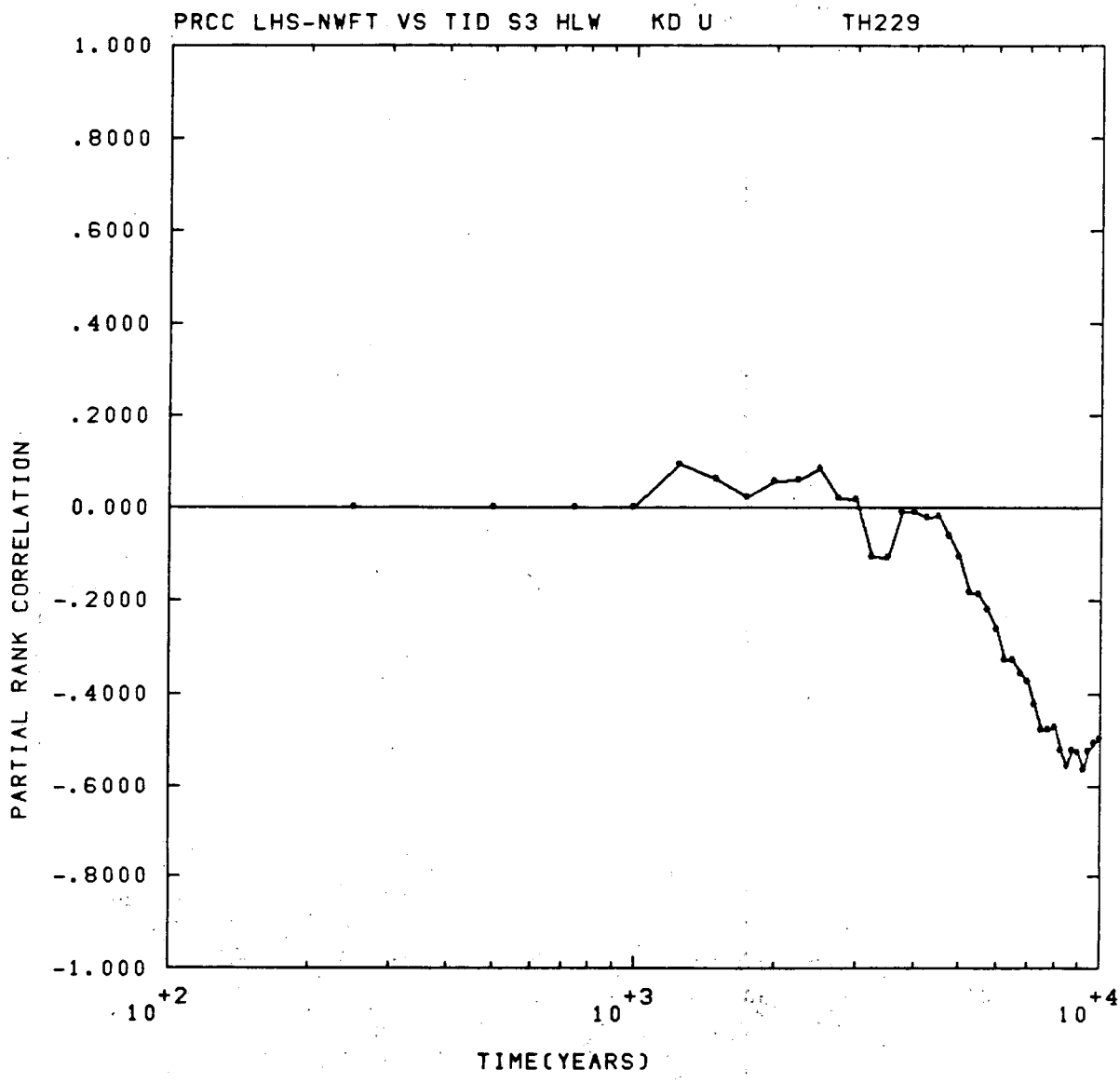

PRCC LHS-NWFT VS TID S3 HLW K UP AQ NP237
1.000
.8000
.6000
.4000
.2000
0.000
-.2000
-.4000
-.6000
-.8000
-1.000
PARTIAL RANK CORRELATION
10+2
10+3
10+4
TIME(YEARS)

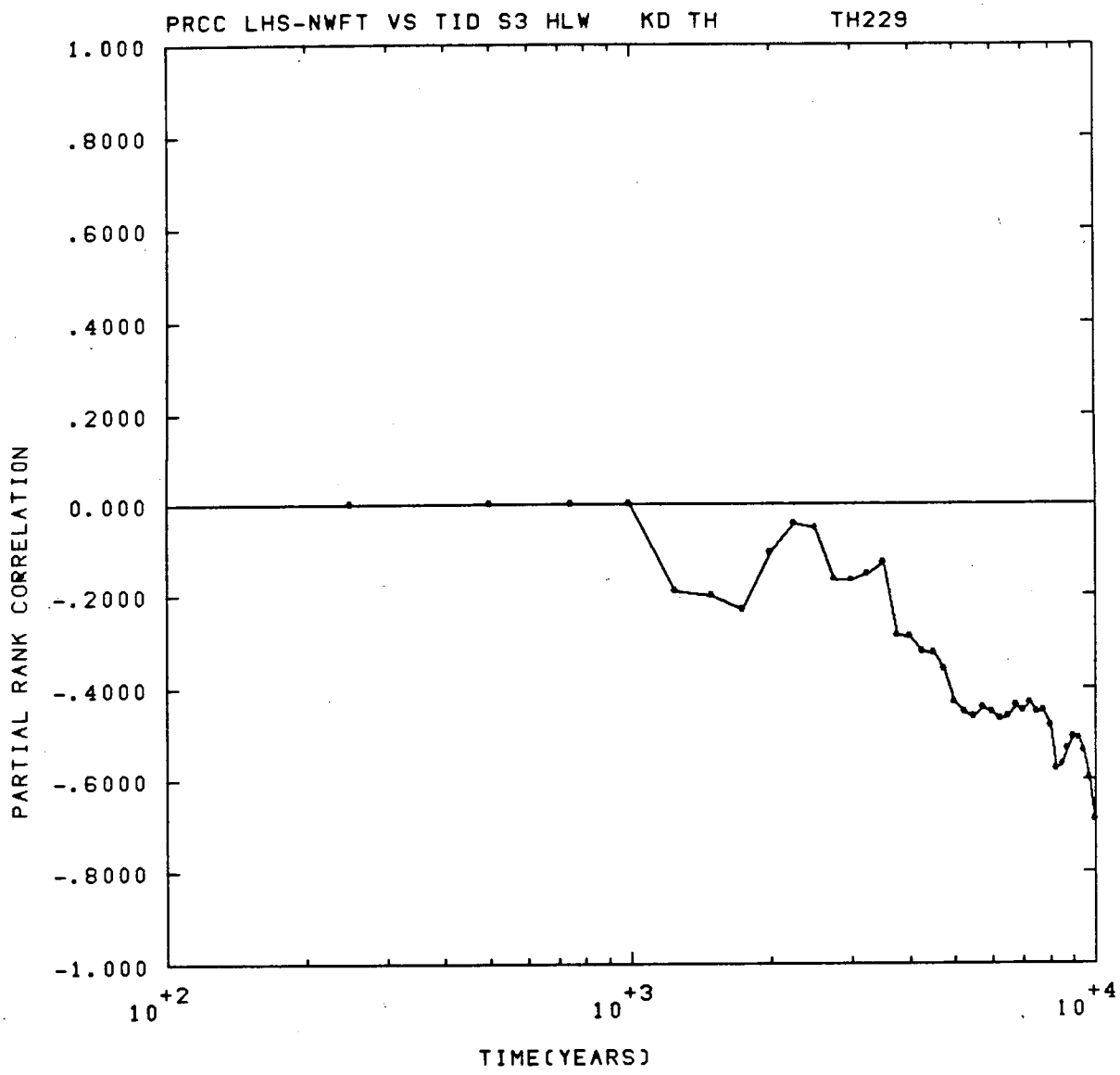

PRCC LHS-NWFT VS TID S3 HLW REL TIME NP237
1.000
.8000
.6000
.4000
.2000
0.000
-.2000
-.4000
-.6000
-.8000
-1.000
PARTIAL RANK CORRELATION
10+2
10+3
10+4
TIME(YEARS)

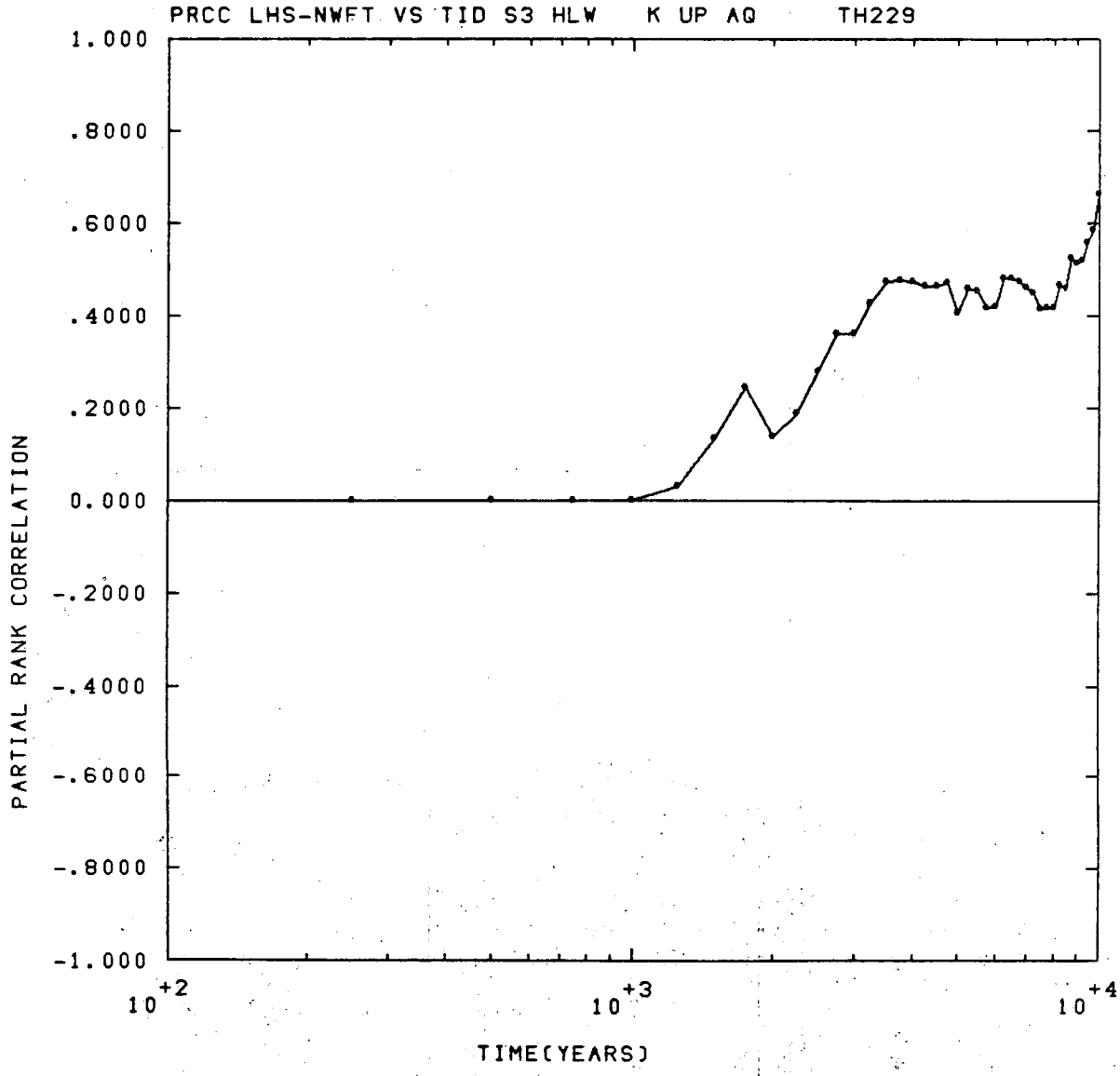

PRCC LHS-NWFT VS TID S3 HLW KD U U 233
1.000
.8000
.6000
.4000
.2000
0.000
-.2000
-.4000
-.6000
-.8000
-1.000
PARTIAL RANK CORRELATION
10+2
10+3
10+4
TIME(YEARS)

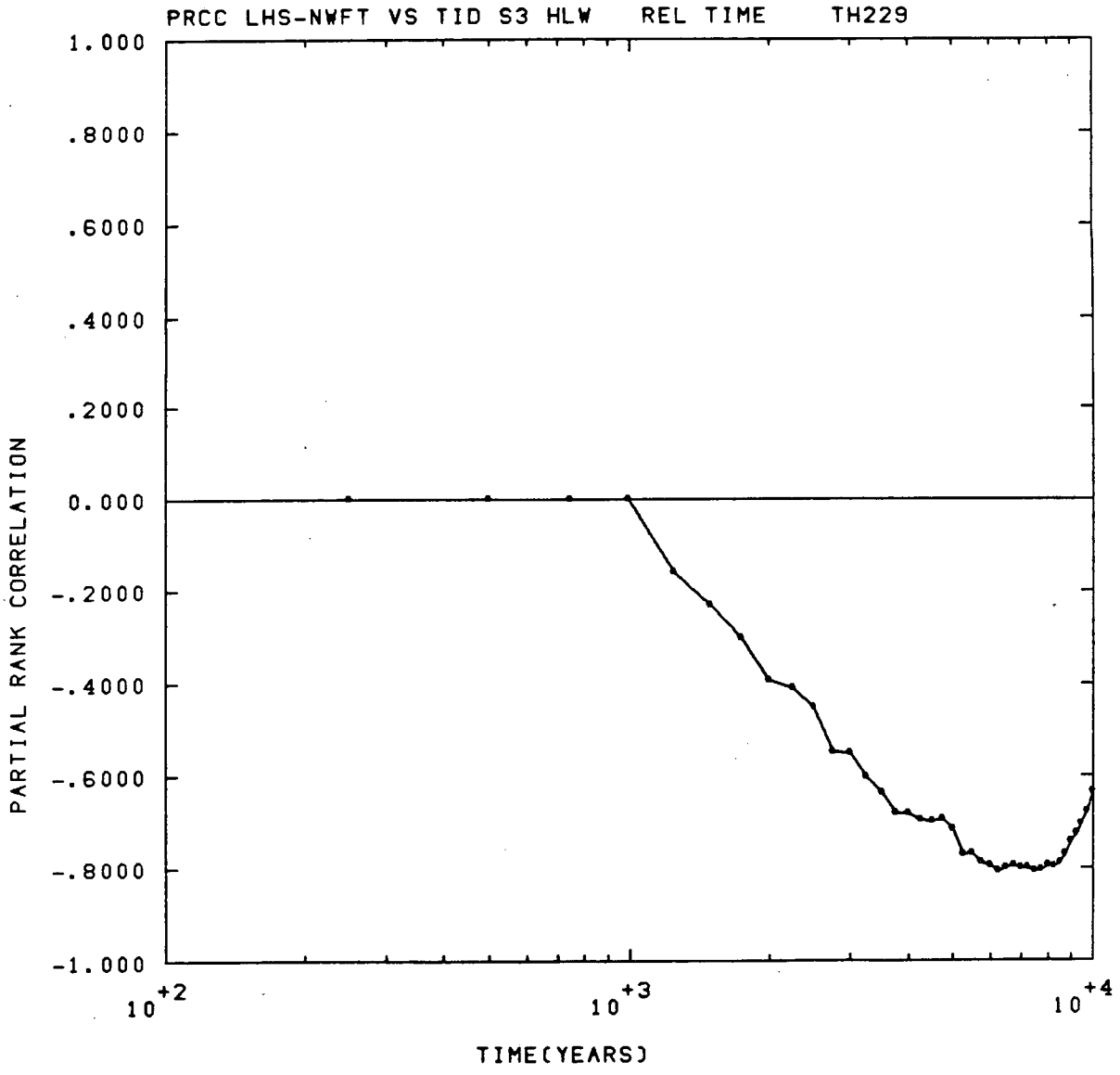

PRCC LHS-NWFT VS TID S3 HLW K UP AQ U 233
1.000
.8000
.6000
.4000
.2000
0.000
-.2000
-.4000
-.6000
-.8000
-1.000
PARTIAL RANK CORRELATION
10+2
10+3
10+4
TIME(YEARS)


PRCC LHS-NWFT VS TID S3 HLW REL TIME U 233
1.000
.8000
.6000
.4000
.2000
0.000
-.2000
-.4000
-.6000
-.8000
-1.000
PARTIAL RANK CORRELATION
10+2
10+3
10+4
TIME(YEARS)


PRCC LHS-NWFT VS TID S3 HLW KD U TH229
1.000
.8000
.6000
.4000
.2000
0.000
-.2000
-.4000
-.6000
-.8000
-1.000
PARTIAL RANK CORRELATION
10+2
10+3
10+4
TIME(YEARS)


PRCC LHS-NWFT VS TID S3 HLW KD TH TH229
1.000
.8000
.6000
.4000
.2000
0.000
-.2000
-.4000
-.6000
-.8000
-1.000
PARTIAL RANK CORRELATION
10+2
10+3
10+4
TIME(YEARS)


PRCC LHS-NWFT VS TID S3 HLW K UP AQ TH229
1.000
.8000
.6000
.4000
.2000
0.000
-.2000
-.4000
-.6000
-.8000
-1.000
PARTIAL RANK CORRELATION
10+2
10+3
10+4
TIME(YEARS)


PRCC LHS-NWFT VS TID S3 HLW REL TIME TH229
1.000
.8000
.6000
.4000
.2000
0.000
-.2000
-.4000
-.6000
-.8000
-1.000
PARTIAL RANK CORRELATION
10+2
10+3
10+4
TIME(YEARS)

STEPWISE REGRESSION ANALYSIS FOR THE PREVIOUS EXAMPLE OF THIS SECTION USING ALL 105 OBSERVATIONS

The remainder of the pages in this section contain the regression analysis results. Page 134 shows some transformations made outside of the regression program to create new variables such as retardation factors. Page 135 shows transformations for variables made within the regression program. The regression parameter cards as described in SAND79-1472 are listed on page 136. The results of the regression analysis on raw data for U233 is given on page 137. Pages 138 to 144 contain the results of regression on ranks for U233.

Summaries of the regression results on raw and rank transformed data are given on pages 145 and 146 respectively. An examination of page 145 shows the analysis on raw data to lack consistency of variable selection from set to set and to give poor fits (low R^2 values). On the other hand the analysis on ranks shown on page 146 does show consistency of variable selection and provides improved fits to the data.

AN EXAMPLE OF A PROGRAM FOR TRANSFORMING SOME VARIABLES TO CREATE NEW VARIABLES OUTSIDE THE REGRESSION PROGRAM - THESE VARIABLES ARE STORED ON DISK 10 PRIOR TO THE EXECUTION OF THE STEPWISE PROGRAM

```
      PROGRAM STAN(INPUT,OUTPUT,TAPE1,TAPE10)
      DIMENSION XIN(20),XOUT(25)
      N=105
      DO 100 I=1,N
      READ(1)XIN
      CO=2.73*(1.0-XIN(13))/XIN(13)
      CE=2.73*(1.0-XIN(15))/XIN(15)
      DO 10 J=1,5
      LO=2*J-1
      LE=2*J
      XOUT(LO)=1.0+XIN(J)*CO
      XOUT(LE)=1.0+XIN(J)*CE
   10 CONTINUE
      XOUT(11)=XIN(6)
      XOUT(12)=XIN(7)
      XOUT(13)=XIN(8)
      XOUT(14)=XIN(9)
      XOUT(15)=XIN(10)
      XOUT(16)=1.0/XIN(11)
      XOUT(17)=XIN(12)
      XOUT(18)=1.0/XIN(13)
      XOUT(19)=XIN(14)
      XOUT(20)=XIN(15)
      XOUT(21)=XIN(16)
      XOUT(22)=XIN(17)
      XOUT(23)=XIN(18)
      XOUT(24)=XIN(19)
      XOUT(25)=XIN(20)
      WRITE(10)XOUT
  100 CONTINUE
      REWIND 10
      END
```

SUBROUTINE FOR MAKING TRANSFORMATIONS
WITHIN THE REGRESSION PROGRAM

```
      SUBROUTINE TRANS(X)
      COMMON/IMAN/NRAW,NTRANS,IDROP,IDUM,IRANK
      DIMENSION X(49)
      DO 1 I=1,22
    1 X(I+25)=X(I)*X(I)
      X(48)=X(17)*X(18)
      X(49)=X(19)*X(20)
      RETURN
      END
```

A LISTING OF THE REGRESSION PARAMETER CARDS AS DESCRIBED IN THE PROGRAM USER's GUIDE

```
SANDIA LABORATORIES <> STEPWISE REGRESSION PROGRAM <> COURTESY OF DEPT. OF STATISTICS - KANSAS STATE UNIVERSITY
TITLE,STEPWISE FOR NRC SHORT COURSE RAW DATA VECT 1-105

DATA,25,24,2.                                                          (STAT CONTROL CARD)

                        INPUT CHECK OF PARAMETERS
                        NUMBER OF VARIABLES READ IN =  25
                        NO. OF TRANSFORMED VARIABLES =  24
                        DATA DISPOSITION IS  2
LABEL(1)=RF A CM,RF S CM,RF A PU,RF S PU,RF A U,RF S U,RF A TH,RF S TH,   (STAT CONTROL CARD)
LABEL(9)=RF A NP,RF S NP,S LIM PU,S LIM U,S LIM TH,S LIM NP,DISP,LEACH T,  (STAT CONTROL CARD)
LABEL(17)=COND AQ,POR AQ,COND S,POR S,REL TIME,NUM RMS,TID NP,TID U,TID TH, (STAT CONTROL CARD)
LABEL(26)=X1SQ,X2SQ,X3SQ,X4SQ,X5SQ,X6SQ,X7SQ,X8SQ,X9SQ,X10SQ,X11SQ,X12SQ,  (STAT CONTROL CARD)
LABEL(38)=X13SQ,X14SQ,X15SQ,X16SQ,X17SQ,X18SQ,X19SQ,X20SQ,X21SQ,X22SQ,     (STAT CONTROL CARD)
LABEL(48)=X17*X18,X19*X20                                              (STAT CONTROL CARD)
OUTPUT,CORR,STEPS                                                      (STAT CONTROL CARD)
STEPWISE,SIGIN=0.05,SIGOUT=0.10                                        (STAT CONTROL CARD)
MODEL,23,24,25=1+3+5+7+9+11+12+13+14+15+16+17+18+19+20+21+22+26+28+30+32+34+  (STAT CONTROL CARD)
 36+37+38+39+40+41+42+43+44+45+46+47+48+49.
PRESS                                                                  (STAT CONTROL CARD)
END OF PARAMETERS                                                      (STAT CONTROL CARD)
```

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

AOV TABLE
ANALYSIS OF REGRESSION FOR VARIABLE 24---FID U
(TABLE 1)

SOURCE	D.F.	SS	MS	F	SIGNIFICANCE
REGRESSION	1	.78330192E-05	.78330192E-05	14.65[illegible]750	.0002
RESIDUAL	103	.55038456E-04	.53435792E-06		
TOTAL	104	.62871885E-04			

R**2 IS .1245[illegible]
INTERCEPT IS -.12143362E-04
STANDARD ERROR OF INTERCEPT IS .808323E-04

VARIABLE NUMBER	VARIABLE NAME	REGRESSION COEFFICIENTS	STANDARDIZED REGRESSION COEFFICIENTS	PARTIAL SSQ	T-TEST VALUES	R**2 DELETES	ALPHA HATS
16	LEACH T.	1.3577349	.352969	.0000	3.8287	0.0000	.0002

UNIQUE SEQUENCE NUMBER FOR THIS ANOVA = 108

PRESS IS .64275E-04

The Analysis on Raw Data Shows Only One Variable (Leach Time) to be Significant.

```
SANDIA LABORATORIES <> STEPWISE REGRESSION PROGRAM <> COURTESY OF DEPT. OF STATISTICS - KANSAS STATE UNIVERSITY
                TITLE,STEPWISE FOR NRC SHORT COURSE RANK TRANSFORMED DATA VECT 1-105

                DATA,25,24,2.                                                    (STAT CONTROL CARD)

                                    INPUT CHECK OF PARAMETERS
                                    NUMBER OF VARIABLES READ IN =  25
                                    NO. OF TRANSFORMED VARIABLES =  24
                                    DATA DISPOSITION IS  2
                OUTPUT,CORR,STEPS                                                (STAT CONTROL CARD)
                STEPWISE,SIGIN=0.05,SIGOUT=0.10                                  (STAT CONTROL CARD)
                MODEL,23,24,25=1+3+5+7+9+11+12+13+14+15+16+17+18+19+20+21+22+26+28+30+32+34+   (STAT CONTROL CARD)
                 36+37+38+39+40+41+42+43+44+45+46+47+48+49.
                PRESS                                                            (STAT CONTROL CARD)
                RANK REGRESSION                                                  (STAT CONTROL CARD)
                END OF PARAMETERS                                                (STAT CONTROL CARD)
```

RANK CORRELATION MATRIX

RF A CM	1	1.0000										
RF A PU	3	.0308	1.0000									
RF A U	5	.0033	.0145	1.0000								
RF A TH	7	.0051	-.0314	.0109	1.0000							
RF A NP	9	.0364	.0635	.0711	.0026	1.0000						
S LIM PU	11	.9263	.0114	.0146	.0199	.0155	1.0000					
S LIM U	12	.0356	-.0111	-.0196	.0147	-.0126	.0492	1.0000				
S LIM TH	13	-.0267	-.0204	-.0513	.0389	-.0091	-.0149	.0020	1.0000			
S LIM NP	14	.0057	.0151	-.0078	.0264	.0153	-.0086	-.0107	.0146	1.0000		
DISP	15	.0192	-.0057	.0417	-.0335	-.0026	.0407	-.0115	-.0536	-.0493	1.0000	
LEACH T	16	.0174	-.0019	-.0241	.0167	-.0181	.0516	-.0255	.0387	.0288	-.0200	1.0000
COND AQ	17	-.0629	-.0804	-.0944	-.0883	-.1409	-.0062	.0305	.0117	-.0103	-.0199	.0049
POR AQ	18	.0938	.0730	.1105	.1047	.2508	-.0200	.0149	-.0514	.0111	.0273	.0342
COND S	19	.0734	.0130	-.0210	-.0273	-.0231	-.0183	-.0575	.0019	-.0084	.0042	.0136
POR S	20	.0380	.0207	-.0199	.0126	.0006	.0219	.0312	-.0874	-.0193	.0324	-.0166
REL TIME	21	-.0393	-.0133	.0517	-.0027	-.0058	-.0439	-.0320	-.0174	.0280	.0401	-.0523
NUM RMS	22	-.0133	-.0017	.0258	.0187	-.0097	-.0104	-.0323	-.0120	-.0237	-.0580	.0474
X1SQ	26	.9688	.0190	-.0242	.0792	-.0213	.0233	.0505	-.0213	.0102	-.0063	.0042
X3SQ	28	.1033	.9688	-.0140	-.0397	.0525	.0080	.0367	-.0030	-.0095	-.0064	.0031
X5SQ	30	.0009	-.0240	.9688	.0211	.1065	-.0124	-.0268	-.0700	-.0106	.0408	-.0327
X7SQ	32	.0117	-.0218	.0049	.9688	.0059	-.0099	.0456	.0353	-.0122	.0061	.0096
X9SQ	34	.0025	.0817	.0874	-.0186	.9688	.0244	-.0238	.0245	.0296	-.0006	-.0518
X11SQ	36	.0229	-.0300	.0098	.0427	.0542	.9688	-.0105	-.0198	-.0134	.0278	.0358
X12SQ	37	-.0075	.0114	-.0449	.0298	.0207	.0499	.9688	.0208	.0062	.0355	-.0116
X13SQ	38	-.0365	-.0237	-.0351	.0464	-.0143	.0015	.0141	.9688	.0129	-.0414	.0778
X14SQ	39	.0382	.0265	.0197	.0566	.0088	-.0038	-.0240	.0019	.9688	-.0286	.0072
X15SQ	40	.0038	-.0216	.0238	-.0075	-.0165	.0479	-.0201	-.0443	-.0694	.9688	-.0240
X16SQ	41	.0316	-.0062	-.0477	.0240	-.0120	.0266	-.0113	.0887	.0792	-.0582	.9688
X17SQ	42	-.0500	-.0382	-.0924	-.1038	-.0983	-.0031	.0164	.0104	.0140	.0159	.0154
X18SQ	43	.1007	.0917	.1532	.1276	.2593	.0224	-.0446	-.0621	.0003	.0277	.0668
X19SQ	44	.0609	.0323	-.0268	-.0575	-.0228	-.0248	-.0179	-.0018	.0209	-.0189	.0102
X20SQ	45	.0579	.0419	-.0121	.0100	-.0045	.0353	.0305	-.0538	-.0038	-.0013	-.0200
X21SQ	46	-.0242	-.0159	.1020	-.0176	-.0353	-.0557	-.0302	-.0291	.0223	.0320	-.0586
X22SQ	47	-.0258	.0102	.0156	-.0438	-.0310	.0081	-.0567	-.0464	-.0038	-.0841	.0577
X17*X18	48	.0649	-.0501	-.0385	.0046	.0424	-.0852	.1544	-.0014	-.0285	-.0013	-.0218
X19*X20	49	.0690	.0512	-.0144	-.0307	-.0193	-.0024	-.0010	-.0181	.0032	-.0047	.0047
TID NP	23	-.0483	-.0896	-.1530	-.0664	-.4711	.0529	.0364	-.0174	-.0531	.0313	.4146
TID U	24	.0369	-.0597	-.5808	.0026	-.2298	.0372	.0626	.0235	-.0412	.0246	.1886
TID TH	25	.0208	-.0277	-.3265	-.4736	-.1573	-.0297	.0868	.0229	-.1319	.1132	.1365

NO.	1	3	5	7	9	11	12	13	14	15	16
NAME	RF A CM	RF A PU	RF A U	RF A TH	RF A NP	S LIM PU	S LIM U	S LIM TH	S LIM NP	DISP	LEACH T

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

CORRELATION MATRIX

COND AQ	17	1.0000										
POR AQ	18	-.6708	1.0000									
COND S	19	.0206	-.0309	1.0000								
POR S	20	.0047	-.0111	.7310	1.0000							
REL TIME	21	-.0342	.0387	-.0069	.0054	1.0000						
NUM RMS	22	-.0195	.0036	-.0036	.0143	.0477	1.0000					
X1SQ	26	-.0167	.0642	.0722	.0257	-.0110	.0012	1.0000				
X3SQ	28	-.0948	.0907	.0425	.0289	-.0176	-.0194	.0835	1.0000			
X5SQ	30	-.0946	.1131	-.0259	-.0062	.0550	.0328	-.0252	-.0536	1.0000		
X7SQ	32	-.0708	.0989	-.0351	.0193	.0054	.0078	.0772	-.0294	.0174	1.0000	
X9SQ	34	-.1348	.2465	-.0260	.0004	.0243	-.0370	-.0597	.0732	.1125	-.0188	1.0000
X11SQ	36	-.0356	-.0183	-.0523	-.0131	-.0055	-.0245	.0290	-.0307	-.0127	.0069	.0542
X12SQ	37	-.0040	.0450	-.0127	.0362	-.0308	-.0332	.0071	.0581	-.0618	.0620	.0067
X13SQ	38	.0664	-.0713	-.0063	-.0876	-.0377	-.0305	-.0236	-.0193	-.0543	.0425	.0103
X14SQ	39	.0207	-.0290	-.0293	-.0147	-.0057	-.0807	.0403	.0007	.0175	.0138	.0178
X15SQ	40	-.0203	.0204	.0090	.0319	.0150	-.0453	-.0243	-.0258	.0358	.0244	-.0220
X16SQ	41	-.0003	.0503	.0203	-.0187	-.0863	.0652	.0237	-.0063	-.0579	.0123	-.0378
X17SQ	42	.9688	-.6450	.0242	.0007	-.0152	.0006	-.0083	-.0450	-.1045	-.1001	-.0927
X18SQ	43	-.6635	.9688	-.0469	-.0111	.0440	.0405	.0677	.1097	.1482	.1224	.2520
X19SQ	44	.0647	-.0435	.9688	.7033	-.0088	.0204	.0587	.0548	-.0384	-.0585	-.0254
X20SQ	45	-.0032	-.0135	.6944	.9688	-.0043	.0617	.0400	.0471	-.0026	.0146	-.0074
X21SQ	46	.0351	.0249	.0123	.0137	.9688	.0574	.0090	-.0264	.1015	-.0102	-.0120
X22SQ	47	-.0010	.0085	.0046	.0287	.0522	.9688	-.0197	-.0048	.0260	-.0519	-.0614
X17*X18	48	.3616	.3209	.0240	.0439	-.0643	-.0624	.0958	-.0556	-.0250	.0298	.0478
X19*X20	49	.0407	-.0354	.9068	.8970	.0021	.0486	.0559	.0687	-.0186	-.0263	-.0164
TID NP	23	.4390	-.3396	.2464	.1619	-.4800	.0344	-.0179	-.0867	-.1670	-.0557	-.4934
TID U	24	.4781	-.3315	.1781	.1316	-.4618	.0109	.0831	-.0473	-.5277	.0212	-.2651
TID TH	25	.5564	-.3605	.2291	.1629	-.3992	-.0806	.0185	-.0131	-.3086	-.4058	-.1834
	NO.	17	18	19	20	21	22	26	28	30	32	34
NAME		COND AQ	POR AQ	COND S	POR S	REL TIME	NUM RMS	X1SQ	X3SQ	X5SQ	X7SQ	X9SQ

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

CORRELATION MATRIX

X11SQ	36	1.0000										
X12SQ	37	-.0042	1.0000									
X13SQ	38	-.0095	.0300	1.0000								
X14SQ	39	.0004	-.0169	-.0003	1.0000							
X15SQ	40	.0275	.0177	-.0357	-.0504	1.0000						
X16SQ	41	.0035	.0076	.1268	.0455	-.0714	1.0000					
X17SQ	42	-.0300	-.0099	.0673	.0445	.0089	.0003	1.0000				
X18SQ	43	.0322	-.0132	-.0792	-.0428	.0132	.0839	-.6199	1.0000			
X19SQ	44	-.0704	.0295	-.0075	-.0089	-.0217	.0185	.0647	-.0645	1.0000		
X20SQ	45	-.0068	.0275	-.0651	-.0062	-.0018	-.0212	-.0138	-.0139	.6833	1.0000	
X21SQ	46	-.0225	-.0369	-.0352	-.0167	.0061	-.0917	.0505	.0286	.0149	.0047	1.0000
X22SQ	47	.0003	-.0585	-.0706	-.0604	-.0713	.0653	.0233	.0415	.0346	.0830	.0648
X17*X18	48	-.1314	.1307	.0256	-.0183	.0054	-.0103	.2644	.2055	.0651	.0438	.0097
X19*X20	49	-.0511	.0246	-.0250	-.0152	-.0054	.0065	.0374	-.0447	.9223	.9020	.0210
TID NP	23	-.0158	.0281	.0458	-.0371	.0631	.4043	.3994	-.3431	.2390	.1322	-.4602
TID U	24	-.0035	.0502	.0575	-.0213	.0594	.1968	.4397	-.3550	.1743	.0967	-.4648
TID TH	25	-.1071	.0747	.0580	-.1298	.1237	.1279	.5271	-.3936	.2576	.1385	-.3787

	NO.	36	37	38	39	40	41	42	43	44	45	46
NAME		X11SQ	X12SQ	X13SQ	X14SQ	X15SQ	X16SQ	X17SQ	X18SQ	X19SQ	X20SQ	X21SQ

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

CORRELATION MATRIX

X22SQ	47	1.0000					
X17*X18	48	-.0467	1.0000				
X19*X20	49	.0687	.0570	1.0000			
TID NP	23	.0400	.1716	.1981	1.0000		
TID U	24	.0173	.2387	.1354	.7671	1.0000	
TID TH	25	-.0512	.2801	.2134	.6766	.7587	1.0000

	NO.	47	48	49	23	24	25
NAME		X22SQ	X17*X18	X19*X20	TID NP	TID U	TID TH

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

AOV TABLE
ANALYSIS OF REGRESSION FOR VARIABLE 24---TID U
(TABLE 1)

SOURCE	D.F.	SS	MS	F	SIGNIFICANCE
REGRESSION	1	31988.185	31988.185	52.436463	.0000
RESIDUAL	103	62833.815	610.03704		
TOTAL	104	94822.000			

R**2 IS .33735
INTERCEPT IS 83.520879
STANDARD ERROR OF INTERCEPT IS 4.85538

VARIABLE NUMBER	VARIABLE NAME	REGRESSION COEFFICIENTS	STANDARDIZED REGRESSION COEFFICIENTS	PARTIAL SSQ	T-TEST VALUES	R**2 DELETES	ALPHA HATS
5	RF A U	-.57586564	-.580818	31988.1848	-7.2413	0.0000	.0000

UNIQUE SEQUENCE NUMBER FOR THIS ANOVA = 118

RANK FIT GIVES A RAW DATA NORMALIZED R**2 = .27533482E-01

COEFFICIENT OF INTERPOLATION = .27082919E-01

PRESS IS 64948.

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

AOV TABLE
ANALYSIS OF REGRESSION FOR VARIABLE 24---TID U
(TABLE 1)

SOURCE	D.F.	SS	MS	F	SIGNIFICANCE
REGRESSION	2	49717.436	24858.718	56.215802	.0000
RESIDUAL	102	45104.564	442.20161		
TOTAL	104	94822.000			

R**2 IS .52432
INTERCEPT IS 105.09772
STANDARD ERROR OF INTERCEPT IS 5.35730

VARIABLE NUMBER	VARIABLE NAME	REGRESSION COEFFICIENTS	STANDARDIZED REGRESSION COEFFICIENTS	PARTIAL SSQ	T-TEST VALUES	R**2 DELETES	ALPHA HATS
5	RF A U	-.55368456	-.558446	29492.4664	-8.1667	.2133	.0000
21	REL TIME	-.42929127	-.432983	17729.2512	-6.3319	.3373	.0000

UNIQUE SEQUENCE NUMBER FOR THIS ANOVA = 119

RANK FIT GIVES A RAW DATA NORMALIZED R**2 = .26972345E-01

COEFFICIENT OF INTERPOLATION = .21997096E-01

PRESS IS 47613.

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

AOV TABLE
ANALYSIS OF REGRESSION FOR VARIABLE 24---TID U
(TABLE 1)

SOURCE	D.F.	SS	MS	F	SIGNIFICANCE
REGRESSION	3	65861.336	21953.779	76.563562	.0000
RESIDUAL	101	28960.664	286.73925		
TOTAL	104	94822.000			

R**2 IS .69458
INTERCEPT IS 80.643474
STANDARD ERROR OF INTERCEPT IS 5.40667

VARIABLE NUMBER	VARIABLE NAME	REGRESSION COEFFICIENTS	STANDARDIZED REGRESSION COEFFICIENTS	PARTIAL SSQ	T-TEST VALUES	R**2 DELETES	ALPHA HATS
5	RF A U	-.51549520	-.519929	25344.1328	-9.4015	.4273	.0000
17	COND AQ	.41111573	.414651	16143.8996	7.5034	.5243	.0000
21	REL TIME	-.41719551	-.420783	16729.6575	-7.6384	.5181	.0000

UNIQUE SEQUENCE NUMBER FOR THIS ANOVA = 120

RANK FIT GIVES A RAW DATA NORMALIZED R**2 = .50748162

COEFFICIENT OF INTERPOLATION = .18358317E-02

PRESS IS 31213.

SUMMARY OF STEPWISE REGRESSION ON RAW DATA

Vectors	NP 237	R^2	U233	R^2	TH229	R^2
1-35	Cond S	.15	Leach T	.18	No. Rms. Rel Time Por S	.33
36-70	$(\text{Por A})^2$ Por A Leach	.66	Leach T S. Lim Np $(\text{Leach T})^2$	.55	Leach T So Lim Np $(\text{Leach T})^2$	.56
71-105	Rel Time	.12	S. Lim Np	.57	$(\text{S. Lim Th})^2$ S. Lim Np S. Lim Th RF A Np	.97
1-105	Rel Time Cond S $(\text{Cond S})^2$ Leach T	.27	Leach T	.12	Por S Rel Time No. Rms	.19

Note the inconsistency of variable selection from one set of runs to the next for this analysis on raw data

SUMMARY OF STEPWISE REGRESSION ON RANKS

Vectors	NP237	R^2	U233	R^2	TH229	R^2
1-35	Cond A Rel Time RF A Np Leach T	.57(.74)	Rel Time Cond A RF A U	.99+(.78)	Cond A RF A TH Rel Time	.94(.83)
36-70	Leach T Rel Time RF A Np Cond A Cond S	.06(.86)	RF A U Rel Time Cond A Por S Leach T	.64(.82)	Cond A Rel Time RF A TH $(RF\ A\ TH)^2$	.89(.79)
71-105	RF A NP Rel Time Leach T Cond A Cond S	.41(.81)	RF A U Rel Time Cond A Cond S $(RF\ A\ U)^2$	.47(.74)	Cond A RF A TH Rel Time RF A Np $(RF\ A\ TH)^2$	.41(.68)
1-105	RF A Np Rel Time Leach T Cond A	.48(.73)	RF A U Rel Time Cond A	.51(.69)	Cond A RF A TH Rel Time RF A U Cond S	.59(.75)

These are the variables selected as important by the stepwise regression analysis on ranks. This selection agrees well with the variables identified as important by the PRCC on pages 114 to 118. Note that the notation RF A used here means retardation factor (RF) in the aquifer and is calculated using the KD values listed with the PRCC. Likewise K UP AQ and Cond A both refer to conductivity in the upper aquifer.

NRC FORM 335 (7-77)

U.S. NUCLEAR REGULATORY COMMISSION

BIBLIOGRAPHIC DATA SHEET

1. REPORT NUMBER *(Assigned by DDC)*
NUREG/CR-2350
SAND81-1978

TITLE AND SUBTITLE *(Add Volume No., if appropriate)*

Sensitivity Analysis Techniques: Self-Teaching Curriculum

2. *(Leave blank)*

3. RECIPIENT'S ACCESSION NO.

7. AUTHOR(S)

R.L. Iman, W.J. Conover

5. DATE REPORT COMPLETED
MONTH: January YEAR: 1982

9. PERFORMING ORGANIZATION NAME AND MAILING ADDRESS *(Include Zip Code)*

Sandia National Laboratories
P.O. Box 5800
Albuquerque, New Mexico 87185

DATE REPORT ISSUED
MONTH: June YEAR: 1982

6. *(Leave blank)*

8. *(Leave blank)*

12. SPONSORING ORGANIZATION NAME AND MAILING ADDRESS *(Include Zip Code)*

U.S. Nuclear Regulatory Commission
Division of Waste Management
Office of Nuclear Material Safety and Safeguards
Washington, D.C. 20555

10. PROJECT/TASK/WORK UNIT NO.

11. CONTRACT NO.
FIN A1158

13. TYPE OF REPORT
Formal Report

PERIOD COVERED *(Inclusive dates)*
October 1980 to January 1982

15. SUPPLEMENTARY NOTES

14. *(Leave blank)*

16. ABSTRACT *(200 words or less)*

This report contains discussions and exercises that illustrate the application of the sensitivity analysis techniques developed at Sandia National Laboratories for the Risk Methodology for Geologic Disposal of Radioactive Waste Project. With this report the user may familiarize himself with the application of the Latin Hypercube Sampling (LHS) program and the Stepwise Regression (STEP) program with the groundwater transport model NWFT/DVM to do sensitivity and uncertainty analyses. The user may require the User's Guides for LHS (SAND 79-1473), STEP (SAND 79-1472), and NWFT/DVM (NUREG/CR-2081) to make full use of this self-teaching curriculum. This report is one of a series of self-teaching curricula prepared under a technology transfer contract for the U.S. Nuclear Regulatory Commission, Office of Nuclear Material Safety and Safeguards.

17. KEY WORDS AND DOCUMENT ANALYSIS

17a. DESCRIPTORS

17b. IDENTIFIERS/OPEN-ENDED TERMS

18. AVAILABILITY STATEMENT
Unlimited

19. SECURITY CLASS *(This report)*
Unclassified

21. NO. OF PAGES

20. SECURITY CLASS *(This page)*
Unclassified

22. PRICE
$

NRC FORM 335 (7-77)