63-2-5

SDC

SP-487

Use of Linearized Nonlinear Regression for Simulations Involving Monte Carlo

3 May 19[illegible]

John E. Walsh

(SP Series)

SP-487

Use of Linearized Nonlinear Regression for Simulations Involving Monte Carlo

John E. Walsh

3 May 1962

SYSTEM DEVELOPMENT CORPORATION, SANTA MONICA, CALIFORNIA

USE OF LINEARIZED NONLINEAR REGRESSION FOR SIMULATIONS INVOLVING MONTE CARLO *

John E. Walsh
System Development Corporation

ABSTRACT

Often, for each simulation situation, a computer run involving Monte Carlo is expensive and time consuming. However, simulation outcomes are usually desired for a very large number of situations. One solution to this difficulty is, from a feasible number of runs, to develop a procedure for estimating the simulation outcome for a large class of situations. This estimation is of a regression nature. The type of regression function used should (1) have substantial curve-fitting flexibility (2) be satisfactorily determined from an acceptable number of simulation runs (3) permit isolation of important combinations of effects, and (4) be computationally manageable. It is also highly desirable that (5) the regression analysis has a satisfactory probability basis, so that the accuracy of the fit from the data can be investigated. The linearized nonlinear regression method presented here satisfies (1)-(4), and some nonprobabilistic curve-fitting procedures for determining a regression function are outlined. This regression method seems to also satisfy (5) and the probability basis developed, along with some estimation results, is presented. Use of the linearized nonlinear regression method has desirable features for optimizations.

*Invited paper presented at the First Joint TIMS-ORSA National Meeting, San Francisco, California, November 8-10, 1961

INTRODUCTION

For the cases considered, a simulation model involving Monte Carlo is programmed for application on a high-speed computer. This mathematical model is completely specified except for the values of some constants which will be termed inputs. Some of the inputs refer to the initial and other side conditions. The other inputs are the values of constants that occur in the functional form for the model. Each possible choice for the values of the inputs represents a simulation situation.

Given the inputs, a computer run of the simulation process yields an observed value for each of the outputs that is considered. These outputs represent the results of this random simulation. Of course, any specified function of the directly recorded outputs is also an output. For the cases considered here, all inputs and outputs are real numbers.

For a fixed set of input values, the observed values for a specified output are a random sample from a univariate statistical population. The properties of these statistical populations (a possibly different population occurs for each set of input values) are the quantities of interest. Estimation of the "value" of a specified output refers to estimation of some suitable "average" property of the statistical population for this output.

Considered over a number of sets of input values, the "average" used is a function of the inputs. In general, this function (referred to as the "average" function) can depend on the totality of sets of input values that receive consideration. However, use of this function to furnish "average" values for other sets of inputs ordinarily should not result in great error when these other sets are of an interpolation nature with respect to the sets on which the "average" function is based. As an aid in the development of a suitable probability model, a special type of "average" function (a generalization of the median concept) is used in this paper.

The problem is to develop a satisfactory method for estimating the "average" function for a specified output (referred to as the output) on the basis of an acceptable number of computer runs. Since statistical variation is involved, this estimation falls under the general heading of regression. However, a combination of desirable features occurs that is seldom present for regression analyses based on data from the usual types of experimental sources. These features include: (a) Any desired number of observed values for the output are obtainable for any specified set of values for the inputs (b) All observed values for the output, irrespective of the values for the inputs, are statistically independent (c) There are no missing value difficulties.

For the linearized nonlinear regression method, the "average" function for estimating the output is represented implicitly. This representation is an equation with one side equal to a sum of unknown constants times completely specified functions of the values of the inputs. The other side equals the value of the (unknown) output plus a sum of unknown constants times completely specified functions of the value of the output. Given values (estimates) for the unknown constants, and the values of the inputs, the estimate of the corresponding output is determined by solution of the resulting equation.

Conditions are imposed to assure that there is a unique solution for the output. As is usual for regression analysis, the completely specified functions are chosen on the basis of technical considerations and judgment. Here, however, a poor selection can often be discovered and improved on the basis of additional simulations.

With suitable choice of the completely specified functions, a considerable amount of curve-fitting flexibility can be obtained without the introduction of very many unknown constants. Since only a moderate number of constants are to be estimated, reasonably stable estimates should be obtainable without an excessively large number of simulation runs. The linearity of the representation implies that important combinations of effects (specified linear functions of the unknown constants) can be isolated without difficulty; also that most of the computations to be performed are of a linear nature and therefore not overly difficult.

Two types of methods are presented for determining linearized nonlinear regression functions (that is, for estimating the unknown constants in a regression function). One method does not have a probabilistic basis and the constants are determined on the basis of curve-fitting criteria. For the other method, approximate median estimates and approximate confidence intervals are obtained for the unknown values of the constants. The nonprobabilistic method is easier to apply and, to some extent, its suitability can be checked by further simulations. The probabilistic method allows direct evaluation to be made of the effects of statistical variation on the estimates for the constants in the regression function.

Any probability model developed should, from practical necessity, be valid for virtually all Monte Carlo simulations and linearized nonlinear regression functions that might reasonably be encountered. The probability model given in this paper has this generality of application. A further development, for which work is now underway, consists in deriving approximate confidence intervals for the sum of the terms which involve the inputs (for any specified set of values for these inputs). These confidence intervals would permit investigation of the over-all effect of statistical variation on the regression function for a stated set of input values.

Consider a measure of effectiveness whose value is a function of the direct outputs of the simulation. For example, this might be the total worth of the targets destroyed for an offense versus defense simulation. This measure of effectiveness is itself an output which depends on the inputs (in an unknown fashion) and can be represented by a linearized nonlinear regression function. Stepwise maximization (or minimization) of this output can be conveniently accomplished by appropriate use of the regression method presented in this paper.

It should, perhaps, be emphasized that the estimation function is only intended for use in situations of an interpolation nature. That is, situations where the set of values for the inputs has an interpolation position with respect to the sets of values used in determining the linearized nonlinear regression function.

The next section contains some discussion, a statement of notation, and a brief general introduction to regression analysis. The following section presents a general statement of the linearized nonlinear regression model. Curve-fitting methods for determining a regression function are considered in the next section. The next to last section contains a statement of the probability model and presents point estimates and confidence intervals for the unknown constants in the regression function. The final section contains an outline of the method used to optimize a measure of effectiveness.

DISCUSSION AND NOTATION

In many senses, regression consists of curve-fitting in the presence of statistical variation. Linear regression, which expresses the "average" value of the output as a linear function of the inputs (or completely specified functions of the inputs) has computational advantages but restricted curve-fitting flexibility. In many cases, poor agreement between observed values of the output and their regression estimates may be due to inability of the linear functional form to provide a good fit rather than a result of large statistical variation. A basic problem is to develop a regression method that has substantial curve-fitting flexibility without substantial changes in the desirable computational properties that occur for linear regression.

Let y represent the output while $x_1,\dots,x_k$ represent the inputs. The data for the i-th of n simulations is represented by

$$(y_i;x_{1i},\dots,x_{ki}), \qquad (i = 1,\dots,n).$$

Then, in general, a regression relation can be expressed in the form

$$y = h(x_1,\dots,x_k;A_1,\dots,A_t),$$

where the function h is completely specified except for the values of the unknown constants $A_1,\dots,A_t$. That is,

$$y_i = h(x_{1i},\dots,x_{ki};A_1,\dots,A_t) + e_i, \qquad (i = 1,\dots,n),$$

where the e_i are random errors and, for the situations considered in this paper, statistically independent.

It is to be emphasized that, in general, each of the e_i could have a substantially different (unknown) probability distribution. Also, the use of a regression function with a large amount of curve-fitting flexibility is helpful in reducing the magnitude of the e_i, since a part of this magnitude is due to inability of the regression function to adequately represent the "fixed parts" of the y_i.

LINEARIZED NONLINEAR REGRESSION MODELS

Let the range of possible values for y be $y_L \leq y \leq y_U$, where $y_L = -\infty$ and $y_U = \infty$ are possibilities. The approach consists in expressing the regression function in a transformed manner. Specifically, let $g_1(y),\ldots,g_s(y)$, usually monotonic for $y_L \leq y \leq y_U$, be specified functions of y while $g_{s+2}(x_1,\ldots,x_k),\ldots,g_t(x_1,\ldots,x_k)$ are specified functions of $x_1,\ldots,x_k$. The regression function is expressed as

$$(1)\quad y + A_1g_1(y)+\ldots+A_sg_s(y) = A_{s+1}+A_{s+2}g_{s+2}(x_1,\ldots,x_k) + \ldots + A_tg_t(x_1,\ldots,x_t),$$

where $A_1,\ldots,A_s$ are restricted by the condition that $y + \ldots + A_sg_s(y)$ is a monotonic function of y for $y_L \leq y \leq y_U$. If $h(x_1,\ldots,x_k;A_1,\ldots,A_t)$ denotes the solution of equation (1) for y in the range $y_L \leq y \leq y_U$,

$$y = h(x_1,\ldots,x_k;A_1,\ldots,A_t)$$

is an alternate way to express the regression function. This alternate expression is used to estimate y for given $(x_1,\ldots,x_k)$.

The functions $g_{s+2},\ldots,g_t$ are selected on technical grounds while $g_1,\ldots,g_s$ are selected partially for curve-fitting generality and partially on technical grounds; none of $g_1(y),\ldots,g_s(y)$ is linear in y. Use of y as one of the linear terms of (1) represents a standardization; it also assures that linear regression can occur as a special case.

The linear form of (1) with respect to $A_1,\ldots,A_t$ allows many types of specified linear combinations of $A_1,\ldots,A_t$ to be isolated by the solution of linear equations. Since y is a monotonic function of the quantity

$$A_{s+1} + A_{s+2}g_{s+2}(x_1,\ldots,x_k) + \ldots + A_t g_t(x_1,\ldots,x_k),$$

isolation of specified linear combinations of $A_{s+1},\ldots,A_t$ are of especially great interest.

Not all types of specified linear combinations of $A_1,\ldots,A_t$ can be isolated by the straightforward procedure of solving linear equations. The values of $A_1,\ldots,A_s$ are restricted by the monotonicity condition on $y + A_1g_1(y) + \ldots + A_sg_s(y)$. However, some number of $A_1,\ldots,A_s$ can be considered to have restricted values. Any specified linear combination of $A_{s+1},\ldots,A_t$ and an unrestricted subset of $A_1,\ldots,A_s$ can be isolated by the solution of linear equations.

CURVE-FITTING RESULTS

Here the problem is to obtain point estimates for $A_1,\ldots,A_t$ on the basis of the $(y_i;x_{1i},\ldots,x_{ki})$ and a criterion that seems to be intuitively satisfactory but for which no specific probability basis is established. Actually, estimation of $A_{s+1},\ldots,A_t$ and a maximum number of unrestricted constants of the set $A_1,\ldots,A_s$ is all that is explicitly considered. For convenience of presentation, $A_1,\ldots,A_{r-1}$ are considered to be restricted while the values of $A_r,\ldots,A_s$ are unrestricted.

The relations that furnish the basis for the estimation are

$$(2)\quad y_i+A_1g_1(y_i)+\ldots+A_sg_s(y_i) = A_{s+1}+A_{s+2}g_{s+2}(x_{1i},\ldots,x_{ki})+\ldots+A_tg_t(x_{1i},\ldots,x_{ki})$$

for $i = 1,\ldots,n$. The first step is to combine these relations so that n new relations are obtained which are linear in $A_r,\ldots,A_t$ and do not contain $A_1,\ldots,A_{r-1}$. This can be accomplished in many reasonable ways by forming n overlapping groups of relations, where each group contains r relations. Then, for each group, a combined linear relation is determined wherein $A_1,\ldots,A_{r-1}$ are eliminated.

Let the relations resulting from these combinations be

$$K_{oi} + K_{ri}A_r + \ldots + K_{si}A_s = K_{(s+1)i}A_{s+1} + \ldots + K_{ti}A_t,$$

$i = 1,\ldots,n$, where the K's have determined values. Then least-squares is used to determine estimates for $A_r,\ldots,A_t$. That is, the estimates are the values such that

$$\sum_{i=1}^{n}(K_{oi} + K_{ri}A_r + \ldots + K_{si}A_s - K_{(s+1)i}A_{s+1} - \ldots - K_{ti}A_t)^2$$

is minimum. There are standard procedures, also computer programs, available for handling this least-squares situation.

Given estimates for $A_r,\ldots,A_t$, determination of suitable estimates for $A_1,\ldots,A_{r-1}$ represents a specialized problem that depends on $g_1(y),\ldots,g_s(y)$. However, in any given case where these functions are strictly monotonic and r is small, solution of this specialized problem should not be overly difficult.

In particular, determining a suitable estimate for A_1 should not be difficult for cases where $r = 2$.

The estimation method outlined here is but one of many possible types. However, it does have the advantages of a reasonable intuitive basis and ease of application.

PROBABILITY RESULTS

The output y_i has a probability distribution but the corresponding inputs $x_{1i},\ldots,x_{ki}$ have fixed values. The y_i are independent but the shape of the distribution of y_i is not assumed to bear any definite relation to the shape of the distribution of y_j for $i \neq j$.

The key feature of the probability model, which allows useful results to be developed for heterogeneous cases, is the method used to define what $A_1,\ldots,A_t$ represent. That is, to specify the type of "average" function that is considered.

By some data manipulations, a few "observations" are constructed that are statistically independent and of the form

$$Y(u;v) = A_v + e(u;v), \qquad (u = 1,\ldots,U),$$

where $e(u;v)$ represents the random error. Let $p(u;v)$ be the value of $P[Y(u;v) \leq A_v]$ and define A_v by the requirement that the arithmetic average of $p(u;v)$ over u is equal to $\frac{1}{2}$. Then an approximate median estimate can be obtained for A_v on the basis of $Y(1;v),\ldots,Y(U;v)$. Also, confidence intervals can be obtained for A_v that have confidence coefficients which are nontrivially bounded from below.

The first step consists in dividing the n relations (2) into mutually exclusive groups of size t (some groups may be of size t + 1). To avoid bias, and not as a part of the underlying probability model, randomization is used to make the subdivision into sets (so that all possible subdivisions are equally likely). Next, separately for each set, by solution of t linear equations in t unknowns, a value is determined for each of $A_1,\ldots,A_t$. For n reasonably large compared to t, this furnishes a number of separate "estimates" for $A_v (v = 1,\ldots,t)$.

Next, for each A_v, its "estimates" are grouped into U classes, where each class contains approximately the same number of "estimates" (some classes may contain one more "estimate" than others). The grouping into classes is the same for each A_v and, to avoid bias, this grouping is determined by randomization. The "observation" $Y(u;v)$ is the arithmetic average of the "estimates" for A_v that occur in the u-th class.

The value chosen for U should not be too large and, for convenience in obtaining approximate median estimates, should be odd. Also U should not be too small (say, $U \geq 5$); otherwise suitable values are not available for the lower bounds for the values of confidence coefficients.

The approximate median estimates and confidence intervals for A_v are based on order statistics of the $Y(u;v)$. Let

$$Y_v[1] \leq Y_v[2] \leq \ldots \leq Y_v[U]$$

be the ordered values of $Y(1;v),\ldots,Y(U;v)$ and suppose that U is odd. Then $Y_v[(U + 1)/2]$ is approximate median estimate of A_v. That is,

$$L \leq P\{Y_v[(U + 1)/2] \leq A_v\} \leq 1 - L,$$

where $L < \frac{1}{2}$ but is not greatly different from $\frac{1}{2}$. Here L = .438 for U = 3, = .421 for U = 5, = .414 for U = 7, = .410 for U = 9, and = .407 for U = 11. These results are obtained on the basis of the material presented in ref. 1.

The following relations define approximate confidence intervals for A_v $(v = 1,\dots,t)$

$$P\{Y_v[u_1] \leq A_v < Y_v[u_2]\} \geq \left(\frac{1}{2}\right)^U \sum_{u=u_1}^{u_2-1} \binom{U}{u},$$

where $u_1 < U/2$, $u_2 > U/2$, $Y_v[0] = -\infty$, and $Y_v[n+1] = \infty$. These results follow directly from ref. 1.

If n is large, the distributions of $Y(1;v),\dots,Y(U;v)$ are approximately continuous and should be at least roughly the same. Then $Y_v[(U+1)/2]$ should be very nearly a median estimate of A_v and the values of the confidence coefficients should nearly equal the lower bound values (see ref. 2 and 3).

As for the curve-fitting case, a complication arises because $A_1,\dots,A_s$ are restricted. Here too, the procedure is to first obtain estimates for a set such as $A_r,\dots,A_t$ and then consider estimation for the restricted set $A_1,\dots,A_{r-1}$. The confidence intervals that are obtainable for each of $A_1,\dots,A_{r-1}$ can be useful in determining the estimates for these unknown constants.

The probability model used is based on the general approach presented in ref. 3. That is, this is a special case where the "median" is the generalized percentage point that is considered.

OPTIMIZATION USE

Here the output considered is a measure of effectiveness that depends (in some unknown way) on $x_1,\dots,x_k$. The problem is to determine a set of values for $x_1,\dots,x_k$ that is optimum (for example, that maximizes this output).

As a preliminary step, the values of s,t are chosen and the functions $g_1,\ldots,g_t$ are decided upon. Then a first approximation to the measure of effectiveness, as a function of the x's, is obtained for sets of values that cover a broad region in the $(x_1,\ldots,x_k)$-space. Examination of this resulting regression function should lead to a substantial decrease in the size of the region in which the optimum $x_1,\ldots,x_k$ can be expected to lie.

Additional observed outputs are obtained for values of $x_1,\ldots,x_k$ in the restricted region and the regression function is redetermined exclusively on the basis of observations pertaining to this region. Examination of this second approximation to the measure of effectiveness as a function of $x_1,\ldots,x_k$ should yield a further decrease in the size of the region where the optimum point should lie. Continuation of this procedure through several stages should ordinarily furnish a good approximation to the optimizing value of $x_1,\ldots,x_k$.

REFERENCES

1. Hoeffding, Wassily, "On the distribution of the number of successes in independent trials," Annals of Mathematical Statistics, Vol. 27 (1956), pp. 713-721

2. Walsh, John E., "Approximate probability values for observed number of successes from statistically independent binomial events with unequal probabilities," Sankhyā, Vol. 15 (1955), pp. 281-290

3. Walsh, John E., "Definition and use of generalized percentage points," Sankhyā, Vol. 21 (1959), pp. 281-288

UNCLASSIFIED

System Development Corporation,
Santa Monica, California
USE OF LINEARIZED NONLINEAR REGRESSION
FOR SIMULATIONS INVOLVING MONTE CARLO.
Scientific rept., SP-487, by J. E. Walsh.
3 May 1962, 14p., references.
Unclassified report

DESCRIPTORS: Equations. Monte Carlo Method. Simulation.

Reports that a computer run for each situation involving Monte Carlo is often expensive and time consuming, but that simulation outcomes are usually desired for a very large number of situations. Presents a linearized nonlinear

UNCLASSIFIED

---

regression method which has substantial curve-fitting flexibility, is satisfactorily determined from a number of simulation runs, and is computationally manageable. Outlines some nonprobabilistic curve-fitting procedures for determining a regression function. Concludes that the use of the linearized nonlinear regression method has desirable features for optimizations.

UNCLASSIFIED

UNCLASSIFIED