THEMIS SIGNAL ANALYSIS STATISTICS RESEARCH PROGRAM

EXTENDED USES OF LINEARIZED NONLINEAR REGRESSION FOR RANDOM-NATURE SIMULATIONS

by

JOHN E. WALSH and G. J. KELLEHER

Technical Report No. 4
Department of Statistics THEMIS Contract

Department of Statistics
Southern Methodist University
Dallas, Texas 75222

THEMIS SIGNAL ANALYSIS STATISTICS RESEARCH PROGRAM

EXTENDED USES OF LINEARIZED NONLINEAR REGRESSION FOR RANDOM-NATURE SIMULATIONS

by

JOHN E. WALSH and G. J. KELLEHER

Technical Report No. 4
Department of Statistics THEMIS Contract

September 5, 1968

Research sponsored by the Office of Naval Research
Contract N00014-68-A-0515
Project NR 042-260

DEPARTMENT OF STATISTICS
Southern Methodist University

EXTENDED USES OF LINEARIZED NONLINEAR REGRESSION
FOR RANDOM-NATURE SIMULATIONS

by

John E. Walsh
Southern Methodist University* and
Dallas, Texas, U. S. A.

G. J. Kelleher
Institute for Defense Analyses
Arlington, Virginia, U. S. A.

ABSTRACT

Linearized nonlinear regression (introduced in ref. 1) has substantial curve-fitting capability, computational simplicity, ability to isolate and investigate effects of interest, etc. A probability model was developed that yields approximate median estimates and confidence intervals for the individual regression coefficients. This model is applicable for random-nature simulations if the simulations are statistically independent. This approach allows the outcomes for wide classes of combinations of values for simulation inputs (that specify the situation simulated) to be estimated from a moderate number of simulations. In this extension of the method, the probability model is slightly changed and approximate results with greater practical utility are developed. Median estimates, confidence intervals, and significance tests are developed for specified linear functions of the regression coefficients that are associated with the simulation inputs. Also, properties of least-squares estimates for specified linear functions of the regression coefficients are examined.

*Partially based on research supported by NASA Grant NGR 44-007-028. Also associated with ONR Contract N00014-68-A-0515.

INTRODUCTION

A random-nature simulation, using a high speed computer, can be identified by the values for a set of inputs (initial conditions, side conditions, values of constants in functional form, etc.). Often, the number of inputs is large and several values can be of interest for each input. Thus, an extremely huge number of combinations of input values can be of interest. However, time and expense permit examination of only a very small fraction of these combinations.

One way of handling this difficulty is to develop a suitable regression function, in terms of the inputs, to estimate the value of the (univariate) output being considered. For example, the output might be a measure of effectiveness for the results of the simulation. Since little is usually known about the probability properties for outputs of simulations, the regression function used should have a probability model that is applicable for virtually any situation that could occur.

The linearized nonlinear regression (LNR) method introduced in ref.1 seems to be suitable. A probability model occurs that is usable when the simulations are statistically independent (virtually always the case). The LNR method is capable of great curve-fitting flexibility and computations are simplified by its linear form in the regression coefficients. Also, isolation of effects is simplified by this form.

Specifically, let $x_1, \ldots, x_k$ represent the inputs (values specified and fixed) while y is the random output being estimated, with $y_L \leq y \leq y_U$ being the possible values. Then, y is the solution of the implicit LNR expression

$$y + A_1 g_1(y) + . . . + A_s g_s(y)$$
$$= A_{s+1} + A_{s+2} g_{s+2}(x_1, . . . ,x_k) + . . . + A_t g_t(x_1, . . . ,x_k), \qquad (1)$$

where $A_1, . . . ,A_s$ are such that the lefthand side is a strictly monotonic function of y for $y_L \le y \le y_U$. The completely specified functions $g_1, . . . ,g_s$ are selected for curve-fitting flexibility and for convenience. The completely specified functions $g_{s+2}, . . . ,g_t$ are chosen on technical grounds.

Procedures are given in ref.1 for investigating each of the A's (individual regression coefficients) separately. However, examination of (1) indicates that, when $A_{s+1}, . . . ,A_t$ are considered, linear functions of the form

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

(can be interpreted as representing effects) are more pertinent for investigation. That is, new "parameters" that are t - s specified linear functions of $A_{s+1}, . . . ,A_t$, and algebraically independent, are investigated (some of these linear functions might be individual regression coefficients). These parameters, rather than $A_{s+1}, . . . ,A_t$, are the quantities defined in terms of the probability model used. For $A_1, . . . ,A_s$, however, the individual regression coefficients are considered.

Sometimes, additional linear functions of $A_{s+1}, . . . ,A_t$ are of interest (for example, some of the individual A's). Fortunately, any other linear function of $A_{s+1}, . . . ,A_t$ can be investigated using the statistics obtained for investigating the t - s parameters. This is not unexpected since given values for the parameters determine the value for any specified linear function of $A_{s+1}, . . . ,A_t$. In particular, estimates for the parameters determine an estimate for any such linear function. Properties of such estimates are examined. Under moderately general circumstances, they are approximate median estimates and are approximately unbiased. Also, approximate equal-tail confidence intervals and significance tests are obtainable for any specified linear function of $A_{s+1}, . . . ,A_t$. A more extensive class of investigative procedures is available when n is sufficiently large.

The definitions of $A_1, . . . ,A_s$ and the t - s parameters are of a slightly different type than was used in ref. 1. The change has the advantage of yielding exact median estimates for $A_1, . . . ,A_s$ and the parameters. Also, determination of properties for estimates for arbitrary linear functions of $A_{s+1}, . . . ,A_t$ is simplified. However, some small disadvantages occur with respect to confidence intervals and significance tests for $A_1, . . . ,A_s$ and the parameters. That is, except possibly for equal-tail intervals and tests, the probability levels are less accurately determined with this modified type of probability model.

A least-squares procedure for estimating most of the regression coefficients(including all of $A_{s+1}, . . . ,A_t$) was given in ref.1. This procedure is examined for the situation of estimating the parameters instead of $A_{s+1}, . . . ,A_t$. It is easily seen that the estimates for $A_{s+1}, . . . ,A_t$ directly provide estimates for the parameters. That is, a parameter is a specified linear function of these A's, and direct substitution of the estimates for the A's into this linear function yields a least-squares estimate of the parameter. This is a direct consequence of the fact that the minimum of a function over a specified coordinate system is the same as the minimum over any other equivalent coordinate system. Moreover, the minimizing values for a coordinate system can be obtained by direct solution from those for any other coordinate system.

Here, $A_1, \ldots, A_t$ constitute one coordinate system and $A_1, \ldots, A_s$ plus the t - s parameters another coordinate system.

The least-squares estimation is considered only on a curve-fitting basis in ref.1. Here, some probability properties are examined. Since, $A_1, \ldots, A_s$ and the parameters are not defined with an expectation basis, the expected value of the lefthand side of (1) is not necessarily equal to the righthand side. Thus, the least-squares estimates are not necessarily unbiased (as they would be, irrespective of the covariance matrix for the observations, if the error term always had zero expectation; for example, see ref. 2). However, if the expectation of the lefthand side minus the righthand side is small compared to its variance, the least-squares estimates should be approximately unbiased in a practical sense.

The next section provides an outline of the least-squares method. The following section contains statements of the new and old probability models; also, exact median estimates, and approximate confidence intervals, are obtained for $A_1, \ldots, A_s$ and the parameters. The final section outlines the justification for, and presents, estimates and confidence intervals for arbitrarily specified linear functions of $A_1, \ldots, A_t$.

LEAST-SQUARES APPROACH

This is essentially the same as in ref.1 and is outlined for convenience of the reader. The data are n statistically independent observations $(y_i; x_{1i}, \ldots, x_{ki})$, $(i = 1, \ldots, n)$. The problem is to estimate $A_{s+1}, \ldots, A_t$ and a maximum number r - 1 of unrestricted constants of the set $A_1, \ldots, A_s$, which are considered to be

$A_r, \ldots, A_s$. The basis is substitution of each $(y_i; x_{1i}, \ldots, x_{ki})$ into (1), yielding n relations in terms of these observations.

First, these relations are combined 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 done by forming n overlapping groups of relations, with each group containing, say, r relations (the extra relation covers the possibility of a degenerate set of linear equations). For each group, a linear combination is taken wherein $A_1, \ldots, A_{r-1}$ are eliminated. The resulting relations are

$$K_{oi} + K_{ri}A_r + \ldots + K_{si}A_s - K_{(s+1)i}A_{s+1} - \ldots - K_{ti}A_t = 0,$$

and $A_r, \ldots, A_t$ are determined as the values that minimize the sum of the squares of these quantities. Given estimates for $A_r, \ldots, A_t$, determination of estimates for $A_1, \ldots, A_{r-1}$ is a specialized problem that depends on $g_1(y), \ldots, g_s(y)$. The parameters are estimated by first estimating $A_{s+1}, \ldots, A_t$ and then substituting these estimates into the linear expressions for the parameters.

PROBABILITY MODELS AND SOME RESULTS

The output y_i has a probability distribution but $x_{1i}, \ldots, x_{ki}$ are fixed. The y_i are independent and can have arbitrarily different distributions. The key feature, which allows useful results to be developed for such heterogeneous cases, is the way of defining each of $B_1, \ldots, B_t$, where $B_1 = A_1, \ldots, B_s = A_s$ and $B_{s+1}, \ldots, B_t$ are the specified linear functions of $A_{s+1}, \ldots, A_t$ that are called the parameters.

A few dummy observations are constructed that are statistically independent and of the form

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

Where U is odd and, ordinarily, $5 \leq U \leq 15$. For fixed v, let $Y_M(v)$ be the median of the $Y(u;v)$, and define $F_v(x)$ by

$$\sum_{u=(u+1)/2}^{U} \binom{U}{u} F_v(x)^u [1 - F_v(x)]^{U-u} \equiv P[Y_M(v) \leq x].$$

Then, $Y_M(v)$ can be considered the median of a random sample of size U from the population with cumulative distribution function (cdf) $F_v(x)$. The constant B_v is defined to be a median of $F_v(x)$, and is virtually always unique.

This definition differs only slightly from that used previously. Let $G_v(x)$ be the arithmetic average of cdf's for the individual $Y(u;v)$. In ref.1, A_v is defined to be a median of $G_v(x)$. However, as shown by use of the expansion in ref.3, $F_v(x) \doteq G_v(x)$, where the approximation is quite close when there is small variation among the cdfs for the individual $Y(u;v)$. As indicated by the way the $Y(u;v)$ are constructed, this variation should be small.

The first construction step consists in dividing the n relations (see the preceding section) into mutually exclusive groups of size t (some may be of size t + 1). To avoid bias and encourage uniformity, but not as part of the probability model, the subdivisions are made by randomization (all possible subdivisions equally likely). Next, separately for each set, a value is determined for each of $B_1, \ldots, B_t$ (by solving t linear equations in t unknowns).

Then, for each B_v, its "estimates" are grouped into U nonoverlapping classes, where the class sizes are approximately the same (some classes may contain one more "estimate" than others). The grouping into classes is the same for all the B_v. To avoid bias and encourage uniformity, this grouping is determined by randomization.

Then, Y(u;v) is the arithmetic average of the "estimates" for B_v that occur in the u-th class.

The statistic $Y_M(v)$ is an exact median estimate of B_v. Let $Y_v(1) \leq . . . \leq Y_v(U)$ be the order statistics of the Y(u;v). Then, the equal-tail confidence interval $[Y_v(u'), Y_v(U + 1 - u'))$ for B_v should have a confidence coefficient of approximately

$$(1/2)^U \sum_{u = u'}^{U + 1-u'} \binom{U}{u}$$

where u' < U/2 and the accuracy tends to increase as u' decreases (based on ref.4). Significance tests for the null hypothesis $B_v = B_v^{(o)}$, completely specified, are obtained from these confidence intervals in the usual manner.

Since $F_v(x)$ only approximately equals $G_v(x)$, a confidence interval of the form $[Y_v(u_1), Y_v(u_2))$, where $u_1 < U/2$, $u_2 > U/2$, $Y_v(0) = - \infty$, $Y_v(n+1) = \infty$, only approximately has the lower bound

$$(1/2)^U \sum_{u = u_1}^{u_2} \binom{U}{u} \qquad (2)$$

for its confidence coefficient. When n is large, the distributions of the Y(u;v) are approximately continuous and should be very nearly the same. Then, the confidence coefficient for any interval of this form should have a value near (2), since $F_v(x)$ nearly equals $G_v(x)$ in the important range for x values and results for $G_v(x)$ imply nearness (ref.3 and 5).

Here too, a complication arises because $B_1, . . . ,B_s$ are restricted. The procedure is to first obtain estimates for a maximum unrestricted set and then consider modification of the estimates obtained for the B's in the restricted set.

Also, for n large, confidence interval results that are applicable to independent symmetrical observations with a central median can be used. As outlined in the next section, the distributions of $Y(1;v), \ldots, Y(n;v)$ tend to symmetry about B_v as n increases. Thus, many of results for investigating the common mean of symmetrical populations (ref.6) are approximately usable.

MATERIAL FOR ARBITRARY LINEAR FUNCTIONS

Consider use of the $Y(u;v)$, $(u = 1, \ldots, U; v = 1, \ldots, t)$, for investigating an arbitrary but specified linear function of $A_1, \ldots, A_t$, which is also a (determined) linear function of $B_1, \ldots, B_t$.

Let us examine a $Y(u;v)$. This is a sum of independent quantities. The Central Limit Theorem indicates that $Y(u;v)$ should have a distribution that is at least roughly symmetrical, especially in the central part. Also, due to symmetrical influence of the randomizations used in their construction, $Y(1;v), \ldots, Y(U;v)$ should have approximately the same expectations (virtually always exist) and at least roughly the same distribution.

The distribution of $Y_M(v)$ should be noticeably more symmetrical than that of any $Y(u;v)$. That is, the important range of x values is concentrated much nearer the central parts of the distributions for the $Y(u;v)$. Also, given that they have a common central value, symmetry of the distributions for the $Y(u;v)$ implies a symmetrical distribution for $Y_M(v)$. The small differences in expectations for the $Y(u;v)$ should not have much effect on the approximate symmetry of $Y_M(v)$. Use of the randomizations tends to symmetrize these differences in $G_v(x)$, which approximately equals $F_v(x)$ for the important range of x values.

Thus, in most cases (virtually all cases when n is sufficiently large), the expectation of $Y_M(v)$ approximately equals B_v. This implies that the expectation of any specified linear function

$$c_1 Y_M(1) + \ldots + c_t Y_M(t) \qquad (3)$$

approximately equals

$$c_1 B_1 + \ldots + c_t B_t, \qquad (4)$$

so that (3) is an approximately unbiased estimate of (4). Also, since (3) is a sum of independent quantities with approximately symmetrical distributions, its distribution is approximately symmetrical. Hence, (3) is also an approximate median estimate of (4). The case of linear functions in $B_{s+1}, \ldots, B_t$ occurs when $c_1 = \ldots = c_s = 0$.

The situation is not so attractive in developing confidence intervals for (4). However, consider the independent statistics

$$Z(u) = c_1 Y(u;1) + \ldots + c_t Y(u;t), \quad (u = 1, \ldots, U).$$

On the basis of the above discussion, the Z(u) should have expectations that roughly equal (4). Also, from the Central Limit Theorem, their distributions should be approximately symmetrical. Thus, the equal-tail confidence interval results in ref.4 (also see ref.6) should be usable with, say, $P[Z(u) \leq (1)]$ not differing from 1/2 by more than .05 and approximate continuity for the distributions of the Z(u). For n large, none of these probabilities should differ much from 1/2, since the distributions of the Y(u;v) tend to symmetry. Then, confidence intervals like those given in the preceding section, using Z(u) in place of Y(u;v), are applicable.

REFERENCES

1. John E. Walsh, "Use of linearized nonlinear regression for simulations involving Monte Carlo," Operations Research, Vol. 11 (1963), pp. 228-235

2. ______________, Handbook of Nonparametric Statistics, III, Van Nostrand, 1968, pp. 19-24

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

4. ______________, "Some bounded significance level properties of the equal-tail sign test," Annals of Math. Stat., Vol. 22 (1951), pp. 408-417

5. ______________, "Definition and use of generalized percentage points," Sankhyā, Vol. 21 (1959), pp. 281-288

6. ______________, Handbook of Nonparametric Statistics, Van Nostrand, 1962, Chapter 7