THEMIS SIGNAL ANALYSIS STATISTICS RESEARCH PROGRAM

DEVELOPMENT AND ANALYSIS OF MODIFIED PAIRED COMPARISONS BY USE OF LINEARIZED NONLINEAR REGRESSION

by

John E. Walsh

Technical Report No. 19
Department of Statistics THEMIS Contract

October 19, 1968

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

DEPARTMENT OF STATISTICS
Southern Methodist University

DEVELOPMENT AND ANALYSIS OF MODIFIED PAIRED COMPARISONS BY USE OF LINEARIZED NONLINEAR REGRESSION

John E. Walsh

Southern Methodist University*

ABSTRACT

The linearized nonlinear regression method (Walsh [1963]) has substantial curve-fitting flexibility. It also permits isolation and probabilistic investigation of pertinent effects, and can be used for developing paired comparisons for persons, animals, or plants ("items" of a given type). Here, an item is identified by the values for specified characteristics, and two kinds of "treatment" (e.g., exposure and nonexposure to radiation) are compared with respect to observed values for a given characteristic. Ideally, two items being compared for treatment effect should be the same with regard to the other characteristics (those not used for comparison). That is, they should have the same set of values for these other characteristics. This ideal situation seldom occurs. However, by suitable use of the linearized nonlinear regression model, composite items can be constructed (for given treatment) that are the same with respect to the other characteristics. Modified paired comparisons are obtained on the basis of these composite items. The probability properties of modified paired comparisons can be very heterogeneous, so that special concepts and statistical techniques are needed. Two approaches for development of tests and confidence intervals are given. Some applications involving exposure to radiation are discussed.

*Research partially supported by NASA Grant NGR 44-007-028. Also associated with ONR Contract N00014-68-A-0515.

1. INTRODUCTION

This paper is concerned with development of generally applicable methods for developing paired comparisons for "items" (persons, animals, or plants) of a given type. The comparison is with respect to the influence of two kinds of "treatment" on a specified univariate characteristic of the items. For example, the items might be people, the treatment could be exposure and nonexposure to nuclear radiation, and the characteristic investigated might be an index of premature aging.

Other characteristics (physical and environmental) besides the one used for comparison can be needed to satisfactorily identify an item. The items receiving one kind of treatment usually differ from nearly all of those receiving the other kind of treatment with respect to the combination of values for these other characteristics. Thus, determination of whether differences between observed values (of the comparison characteristic) are actually due to different treatment is difficult. That is, the disagreement could be due to different combination of values for the other characteristics, rather than different treatment. This is further complicated by the presence of statistical variation. Some approach that investigates treatment effects in a probabilistic manner (accounting for statistical variation) is needed.

Often, specialized information about the probabilistic properties of the observed characteristic for comparison is lacking. Then, the statistical method needs to be of a general nature. The linearized nonlinear regression model that is outlined in section 3 (see Walsh [1963] for a more detailed statement) seems to provide a satisfactory basis for many situations. For example, a modification of comparisons involving pairs of items can be developed (different treatments for the items of a pair). Instead of using actual items, two composite items are constructed

for each pair (with respect to observed values of the characteristic for comparison). The composite items of a pair are constructed to be the same with respect to the combination of values for the other characteristics. Thus, if the model is capable of approximating the true situation, the difference in observed values for a modified pair can be attributed (approximately) to treatment difference. Then, by development of suitable statistical techniques for analysis of modified paired comparison data, treatment, treatment difference can be investigated.

Section 2 contains a general discussion of the regression approach. This is followed by an outline of the linearized nonlinear regression model. The associated probability model is outlined in section 4. Two ways of constructing composite items are given in section 5. Two analysis methods for modified paired comparisons are outlined in section 6. Finally, some potential uses of these results, in particular to investigating exposure to radiation, are discussed.

2. GENERAL DISCUSSION OF REGRESSION

Each item can be represented by a "vector" (multidimensional point). The first coordinate of this vector corresponds to the characteristic being investigated while the other coordinates correspond to the physical and environmental characteristics that are believed to have an important influence on the probability distribution of the first coordinate. Each characteristic has a set of possible values (or levels) and the value of the vector for an item is obtained by giving each coordinate the value of the corresponding characteristic. Notationally, a vector is of the form $(y;x_1, \ldots, x_k)$, where y corresponds to the characteristic being investigated and the x's correspond to the k other characteristics.

Let y_i , x_{1i} ,..., x_{ki} be the values of these characteristics that occur for the i-th item of the n items for which data are available. Then, the i-th item is represented by $(y_i; x_{1i}, \ldots, x_{ki})$.

For the situations considered, the values of the x's are assumed to be fixed quantities and the value of y is considered to be a random variable. A regression approach is convenient for investigating y in such a manner that the influence of the x's receives explicit consideration. Stated in a general manner, y is expressed in the form

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

where the function h is completely specified except for the values of the "regression coefficients" A_1 ,..., A_t and e' is a random variable. A "true" set of values exists for the A's, but these values are virtually always unknown. Thus, the A's occur as unknown parameters. Ideally, h should be chosen so that the general magnitude of e' is as small as possible. The function $h(x_1, \ldots, x_k; A_1, \ldots, A_t)$ is called the regression function of y on x_1 ,..., x_k. Using regression terminology, y is the dependent variable and x_1 ,..., x_k are the independent variables for the regression.

Using this representation, the observed value of y for the i-th item is expressed in the form

$$y_i = h(x_{1i}, \ldots, x_{ki}; A_1, \ldots, A_t) + e_i .$$

It is often reasonable to assume that the e_i are statistically independent. However, for heterogeneous situations, this is one of the few assumptions that is acceptable. There is seldom much justification for the assumptions that the e_i have zero expected values, that they are sample values from the same population, that they have the same variance, etc.

The lack of specialized information about the probability properties of the e_i is one of the fundamental difficulties encountered in analyzing heterogeneous data. Other difficulties occur with respect to selection of the function h. If an elementary form, such as linear, is used for h, isolation of effects due to the independent variables is rather easily accomplished. For example, composite items for paired comparisons can be constructed without substantial computational effort. On the other hand, use of an unsuitable form for h can result in poor agreement between y_i and $h(x_{1i}, \ldots, x_{ki}; A_1, \ldots, A_t)$ over the values of i. In many cases, the failure of a regression function to provide a good fit to the data may be largely due to the restricted curve-fitting capability of the functional form that is used for h. A basic problem in obtaining regression functions that are suitable for heterogeneous data is to greatly increase the curve-fitting capability of the regression function without substantial changes in the desirable manipulation-computation properties that occur for elementary forms such as linear. The linearized nonlinear regression model seems to furnish a satisfactory solution to this problem.

3. LINEARIZED NONLINEAR REGRESSION MODEL

Let the range of possible values for y be $y_L \le y \le y_U$. The approach for linearized nonlinear regression consists in expressing the regression function in a transformed manner. Specifically, let $g_1(y), \ldots, g_s(y)$ 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 $x_1, \ldots, x_k$. Then, y is implicitly expressed in

the form

$$(1) \quad y+A_1g_1(y) + \ldots + A_sg_s(y)$$

$$= A_{s+1} + A_{s+2}g_{s+2}(x_1, \ldots, x_k) + \ldots + A_tg_t(x_1, \ldots, x_k)+e'' ,$$

where $A_1, \ldots, A_s$ are such that the lefthand side of (1) is a monotonic function of y for $y_L \le y \le y_U$. Let $h(x_1, \ldots, x_k; A_1, \ldots, A_t)$ be the solution of (1) for y when $e'' = 0$. Then, expression (1) is equivalent to an expression of the form

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

where the function h can have a substantial amount of curve-fitting capability. The form of (1) allows linear manipulations to be used in isolating effects that are expressed in terms of the A's. Since y is a monotonic function of

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

when statistical variation is neglected, isolation of specified linear combinations of $A_{s+1}, \ldots, A_t$ is of special interest. Although the restriction imposed on $A_1, \ldots, A_s$ places some limitation on the use of linear manipulations, the computational aspects of the linearized non-linear model seem to be at a manageable level.

The primary purpose of $g_1(y), \ldots, g_s(y)$ is to furnish sufficient curve-fitting capability. If y_L and y_U are finite, this can often be accomplished by letting $s = 2$, $g_1(y) = y^2$, and $g_2(y) = y^3$. The choice of $s = 3$, $g_1(y) = y^2$, $g_s(y) = y^3$, and $g_3(y) = y^4$ is available if greater curve-fitting capability is desired. It is anticipated that finite values can be used for y_L and y_U in nearly all cases.

4. ASSOCIATED PROBABILITY MODEL

The dependent variable y_i for the i-th item is considered to have a probability distribution but the independent variables $x_{1i}, \ldots, x_{ki}$ are considered to be fixed. The y_i are assumed to be statistically independent (i.e., the e_i are independent) but the shape of the distribution for y_i does not necessarily have any relation to the shape of the distribution of y_j if $i \neq j$.

The key feature of the probability model is the procedure used to define what the parameters $A_1, \ldots, A_t$ represent. These definitions are intuitively meaningful and also allow useful probability results to be developed for heterogeneous situations. This procedure is a generalization of the median estimation concept.

Suppose that the total number n of items is not too small. By some data manipulations (see Walsh [1963]), a few "observations" Y(u,v) can be constructed that are independent, approximately continuous, and of the form

$$Y(u;v) = A_v + e(u;v) ,$$

where e(u;v) is a random variable. Let p(u;v) be the (unknown) value of $P[Y(u;v) \leq A_v]$. Then, the (unknown) value of A_v is defined by the requirement that the arithmetic average of the p(u;v) over u is equal to ½. Now, by the methods given in Walsh [1963], an approximate median estimate and approximate confidence intervals can be obtained for the true but unknown value of $A_v (v = 1, \ldots, t)$.

5. CONSTRUCTION OF COMPOSITE ITEMS

In a general way, let us consider a couple of procedures that might be used to construct the composite items for a pair that is used in a modified paired comparisons investigation of whether treatment effects

are statistically significant. Under the null hypothesis, treatment has no effect and the same linearized nonlinear regression model can be used for all items.

First, the values of $x_1, \ldots, x_k$ that are to occur for this pair are specified. Then, for a given treatment, one problem is to construct an item with these values for the x's. That is, a suitable subset of the data with this treatment is manipulated to determine an observed value of y that, according to the linearized nonlinear regression model, corresponds to an item with these values for the x's. This value is used as if it were observed for an item with the given treatment with these x's. The same procedure is used to determine the observed value for an item with these values for the x's and the other kind of treatment. These two y values then constitute the observations for this pair. Here, nonoverlapping subsets of data are used for the various pairs that are developed. Thus, a disadvantage of the construction of composite items is that the data for a number of items is needed in order to construct one composite item.

One way of determining an observed value for y is to use a suitable subset of the data to individually estimate $A_1, \ldots, A_s$ and to estimate the value of

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

Using these estimates as if they were the true values, solution of

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

determines a corresponding observed value for y.

A second way of determing an observed value for y is to linearly combine the data on t + 1 items so that the weighted sum of the individual regression expressions is of the form that occurs when the specified values $x_1, \ldots, x_k$ occur. That is, weights w_i (not necessarily positive) are determined so that

$$\sum_{i=1}^{t+1} w_i[y_i + A_1 g_1(y_i) + \ldots + A_s g_s(y_i)]$$
$$= \sum_{i=1}^{t+1} w_i y_i + A_1 g_1 \left(\sum_{i=1}^{t+1} w_i y_i\right) + \ldots + A_s g_s\left(\sum_{i=1}^{t+1} w_i y_i\right)$$

and

$$\sum_{i=1}^{t+1} w_i = 1, \qquad \sum_{i=1}^{t+1} w_i g_r(x_{1i}, \ldots, x_{ki}) = g_r(x_1, \ldots, x_k),$$

$$(r = s + 2, \ldots, t).$$

Then, $\sum_{i=1}^{t+1} w_i y_i$ is the observed value for y.

6. ANALYSIS OF MODIFIED PAIRED COMPARISONS

Having constructed the modified paired comparisons, the difference of the values for each pair can be formed. More generally, each value of a pair can be transformed by use of the same function and the difference of the two transformed values can be formed. These differences can be used to compare the two treatments (with respect to the specified characteristic).

Due to the rather general ways in which the modified pairs are constructed, special statistical procedures will be needed for the analysis

of the differences. Even though the composite items of a pair have the same values for $x_1, \ldots, x_k$, the observed values for y do not necessarily have the same distribution under the null hypothesis. Hence, a difference is not necessarily symmetrically distributed about zero under the null hypothesis. Also, the distribution for one difference may be greatly different from that for another difference. In spite of these difficulties, development of statistical techniques that are satisfactory for analyzing these independent differences is possible.

One method, which seems especially suitable when the second way is used to construct composite items, is to investigate the value of a special parameter which is defined as follows:

> For each difference, consider the (unknown) probability that the difference is less than the parameter. The (unknown) value of the parameter is determined by the requirement that the average of these probabilities is ½. An approximate median estimate and approximate confidence intervals can be developed for the value of this parameter. In particular, procedures can be developed for testing whether this parameter is zero, which seems to be a reasonable choice for its null value (in most cases).

Another method is based on Walsh [1951] (also see Walsh [1962]). Here, a suitably chosen function of the medians of the distributions for the differences is investigated (for example, the arithmetic average of these population medians). By suitable rise of the material in Walsh [1951], approximate equal-tail confidence intervals can be obtained for the function of medians that is investigated. These confidence intervals yield two-sided significance tests of the null value for the function of the medians. Usually, zero seems to be a suitable choice for this null value.

7. DISCUSSION OF USES

The material of this paper is oriented toward situations where the items can be of a very heterogeneous nature. Consequently, it is applicable to many biological and medical areas. However, the wide applicability of the results indicates that they are most appropriate when a large amount of data is available. That is, the approach is somewhat coarse and mainly useful when a large number of items are available for the investigation. Thus, this method is especially suitable for retrospective studies where information has been obtained on a very large number of items.

One application area, which motivated development of these results, is investigation of exposure of persons to nuclear radiation. This method should be satisfactory for analyzing data collected for persons exposed to radiation in the atomic bomb attacks in Hiroshima and Nagasaki during World War II. It should also be somewhat suitable for analyzing the data from the accidental radiation of the Rongelap people (e.g., see Conard et al [1963]).

REFERENCES

Robert A. Conard et al [1963], Medical survey of Rongelap people eight years after exposure to fallout, Report BNL 780 (T-296), Brookhaven National Laboratory.

John E. Walsh [1951], "Some bounded significance level properties of the equal-tail sign test," Annals of Mathematical Statistics, Vol. 22, pp. 408-417.

John E. Walsh [1962], HANDBOOK OF NONPARAMETRIC STATISTICS: Investigation of Randomness, Moments, Percentiles, and Distributions, D. Van Nostrand Co., pp. 169-170.

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