Computer Programs in Biomedicine 8 (1978) 197–207
© Elsevier/North-Holland Biomedical Press

Section I: Programs, program packages and systems

A COMPUTER PROGRAM FOR MULTIVARIATE RATIO ANALYSIS (MISCAT)

William M. STANISH and Gary G. KOCH
Department of Biostatistics, University of North Carolina, Chapel Hill, NC 27514

and

J. Richard LANDIS
Department of Biostatistics, University of Michigan, Ann Arbor, MI 48109, USA

Analysts must deal frequently with missing data in multivariate analysis. In such cases, estimating the covariance maxtrix V of the dependent variables usually involves initial estimation and iterative adjustment of imputed missing data values, and/or smoothing of an estimate $\hat{V}$ which is not necessarily positive semi-definite. This paper presents an alternative procedure for computing estimates of relevant multivariate parameters in situations where missing data occur at random and with small probability. MISCAT is a computer program which computes multivariate ratio estimates of the means and a corresponding positive semi-definite estimate of the covariance matrix. It is an extension of GENCAT, which is a program for the generalized least squares analysis of categorical data. Thus, one advantage of dealing with missing data in this manner is that variation among the ratio estimates may be conveniently analyzed within MISCAT using asymptotic regression methodology, provided that sample sizes are sufficiently large. An example is given to illustrate such analysis for longitudinal data from a multicenter clinical trial.

Missing data   Multivariate ratio analysis   Asymptotic regression methodology   GENCAT

1. Introduction

The analysis of multivariate data is frequently complicated by missing observations for some of the data vectors. Since interest generally lies in inferences about functions of the mean vector $\boldsymbol{\mu}$ and/or the covariance matrix V, numerous techniques have been suggested for estimating these parameters when there are missing data. Some of them yield an estimate of the covariance matrix which is not, in general, positive semi-definite. Moreover, the various methods of computation typically require estimation of the missing data values and this introduces an additional source of difficulty. For a brief review of the pertinent literature on this topic, see Stanish, Gillings, and Koch [6].

This paper describes both a methodology for the analysis of incomplete multivariate data, and the computer program, MISCAT, which implements the calculations. The methodology is concerned with inference on $\boldsymbol{\mu}$ when $\boldsymbol{y} = (y_1, ..., y_r)'$ is a vector of dependent random variables and some of the observed vectors involve missing data. It is assumed that the missing data occur at random, by which we mean that the observance or non-observance of a dependent variable is unrelated to the value of the variable that would have been observed in the complete data case. The methodology involves multivariate ratio estimation of $\boldsymbol{\mu}$ via indicator random variables which denote the presence or absence of data. The primary advantages of this procedure are 4 fold:

(i) The estimate of the covariance matrix of $\hat{\boldsymbol{\mu}}$ is always positive semi-definite.

(ii) No estimation of missing data values is required.

(iii) There are no distributional assumptions required about the random vector $\boldsymbol{y}$.

(iv) There are no iterations required to obtain the final estimates.

The proposed methodology is asymptotic, and thus appropriate only when the sample size is large (at least 25) within each subpopulation (or domain) of interest. This provides for approximate normality of the ratio estimators and consistency of their estimated covariance matrix, which is based on a linear Taylor series. Upon obtaining the estimators, one may investigate variation among the elements of $\boldsymbol{\mu}$ by fitting linear regression models by weighted least squares and using generalized Wald [7] statistics for hypothesis testing, as described n.ore fully by Koch et al. [4].

In this regard, the computational framework for analyzing the data via asymptotic regression models is the same as that used for the generalized chi-square analysis of categorical data by weighted least squares. Thus, the computer program, MISCAT, was written as an extension of the categorical data analysis program, GENCAT [5]. The extensions involve 2 major features:

1. MISCAT provides for the direct calculation of means, whereas in GENCAT, the user is required to specify a number of functions and a corresponding linear combination of these in order to obtain a single mean score. Furthermore, MISCAT has the additional capability of handling continuous data.
2. When some of the data are missing, the calculation of the means and the corresponding covariance matrix is modified according to the methodology presented in this paper so as to reflect the inherent structure of the missing data.

Otherwise, it should be noted that MISCAT has all the capabilities of GENCAT, together with the extra option to deal with missing data and the improved features for computation of means.

2. Methodology

Let $\boldsymbol{y}' = (y_1, y_2, ..., y_r)$ be a random vector with expectation $\boldsymbol{\mu}' = (\mu_1, \mu_2, ..., \mu_r)$. Suppose that a sample of n observation vectors is taken from a population, and some of the vectors involve missing data. The problem is to find an estimator $\hat{\boldsymbol{\mu}}$ and its covariance matrix under these conditions. One method of estimation which is convenient for missing data situations is based on a generalization of a method suggested by Cornfield [2]. With this method it is assumed that there is an additional random vector $\boldsymbol{u}' = (u_1, u_2, ..., u_r)$ which controls the missing data process, and is such that u_j is independent of y_j, for each j. The component u_j takes the value 1 if the j-th dependent variable is observed, and the value 0 otherwise; thus, the observed random variables are $f_j = y_j u_j$, for $j = 1, 2, ..., r$, where 'missing' is assigned the value zero. The mean μ_j of the j-th dependent variable is estimated by the mean of the data present, i.e., the ratio estimator:

$$R_j = (\sum_{l=1}^{n} f_{jl}/n)/(\sum_{l=1}^{n} u_{jl}/n)$$
$$= \exp\{\log_e(\bar{f}_j) - \log_e(\bar{u}_j)\} \tag{2.1}$$

where l indexes subjects. By construction, this estimate is the same as the mean of the data which are present and is thereby equivalent to the estimate which would be obtained if missing data were replaced by this mean. However, the ratio formulation (2.1) provides its inherent structure relative to the sample as a whole and hence is the basis for the estimation of the corresponding covariance matrix.

When the pertinent random variables are included in one vector $\boldsymbol{g}' = (f_1, f_2, ..., f_r, u_1, u_2, ..., u_r)$, the multivariate ratio estimator of $\boldsymbol{\mu}$ can be expressed as:

$$\boldsymbol{R} = \exp\{\boldsymbol{A} \ \mathbf{log}_e(\bar{\boldsymbol{g}})\} \tag{2.2}$$

where $\bar{\boldsymbol{g}}$ is the sample mean vector of the $\boldsymbol{g}$'s, $\mathbf{log}(\boldsymbol{a})$ transforms each element of $\boldsymbol{a}$ to its natural logarithm, $\mathbf{exp}(\boldsymbol{a})$ transforms each element of $\boldsymbol{a}$ to its antilogarithm, and:

$$\boldsymbol{A} = [\boldsymbol{I}_r, -\boldsymbol{I}_r] \tag{2.3}$$

To estimate the covariance matrix of $\boldsymbol{R}$, let $\boldsymbol{V}_{\bar{g}}$ denote the multinomial model maximum likelihood estimate of the covariance matrix of $\bar{\boldsymbol{g}}$ as discussed by Koch et al. ([4] appendix 2). That is:

$$\boldsymbol{V}_{\bar{g}} = \{V_{\bar{g}_{kk'}}\}$$
$$= \{\sum_{l=1}^{n} (g_{kl} - \bar{g}_k)(g_{k'l} - \bar{g}_{k'})/n^2\} \tag{2.4}$$

Suppose $\boldsymbol{F}$ is a vector of functions of $\bar{\boldsymbol{g}}$. Then a consistent estimator for the covariance matrix of $\boldsymbol{F}$ is:

$$\boldsymbol{V_F} = \boldsymbol{H V_{\bar{g}} H'} \tag{2.5}$$

where $\boldsymbol{H} = [\mathbf{d}\boldsymbol{F}(\boldsymbol{x})/\mathbf{d}\boldsymbol{x} | \boldsymbol{x} = \bar{\boldsymbol{g}}]$ is the matrix of first derivatives of the functions $\boldsymbol{F}$ evaluated at $\bar{\boldsymbol{g}}$. Successive applications of the operations in (2.5) to the compound function (2.2) yield the result that a consistent estimator of the covariance matrix of $\boldsymbol{R}$ is:

$$V_R = D_R A D_{\bar{g}}^{-1} V_{\bar{g}} D_{\bar{g}}^{-1} A' D_R \tag{2.6}$$

where D_a is a diagonal matrix with elements of $\boldsymbol{a}$ on the main diagonal. Furthermore, it is clear from the construction process that this estimated covariance matrix is positive semi-definite.

Suppose there is a set of s subpopulations under investigation, indexed by $i = 1, ..., s$. Then, for each subpopulation, a sample mean vector $\bar{g}_i$ and its corresponding estimated covariance matrix V_i can be constructed in a manner analogous to that described for $\bar{g}$ and $V_{\bar{g}}$ in the previous discussion. To analyze the subpopulations simultaneously, the vectors $\bar{g}_i$ are concatenated to form $\bar{g}$, and the matrices V_i become the diagonal elements of the block diagonal matrix $V_{\bar{g}}$. Equations (2.2) and (2.6) are still applicable for computing $\boldsymbol{R}$ and its estimated covariance matrix, except that in this case:

$$A = [I_r, -I_r] \otimes I_s \tag{2.7}$$

where $\otimes$ denotes Kronecker product and I_k is the identity matrix of rank k.

All of these estimates may be obtained using GENCAT provided that the response data are categorical and the user specifies the appropriate indicator functions and transformations. However, the estimates are automatically computed by MISCAT when the user specifies the dependent variables of interest and the missing data option. Furthermore, MISCAT is capable of analyzing continuous response data as well as categorical data.

Variation among the elements of $\boldsymbol{R}$ may be analyzed with MISCAT via asymptotic regression models [4] of the form:

$$E_A(R) = \boldsymbol{\mu} = X\boldsymbol{\beta} \tag{2.8}$$

where E_A denotes asymptotic expectation, $\boldsymbol{\beta}$ is a vector of unknown parameters to be estimated, and $\boldsymbol{X}$ is a design matrix directed at the relationships among the components of $\boldsymbol{\mu}$ with respect to the independent variables of interest. Model goodness of fit and hypotheses concerning linear combinations of the parameters are tested with Q statistics which have approximate chi-square distributions provided that the sample sizes from the respective subpopulations are sufficiently large than $\boldsymbol{R}$ is approximately multivariate normal. Thus, these aspects of MISCAT are the same as those which are incorporated in GENCAT.

The use of multivariate ratio estimation in missing data methodology deserves some further comment. Ratio estimators have long been used in sampling methodology [1,3]. There, the properties of ratio estimators are satisfactory when the sample size is large since, in that case, their bias is negligible, and they are approximately normally distributed. Large samples also enhance the consistency of the estimate of the covariance matrix of $\boldsymbol{R}$ since that estimate is based on a linearized Taylor series.

The application of the methodology in this paper is not generally recommended in situations where more than ten percent missing data occur for any of the variables under study because the fundamental missing data assumption is more likely to be violated with larger amounts of missing data. This issue can be recognized by recalling the fundamental assumption of the multivariate ratio analysis, namely, that the missing data random variable u_j is independent of the study variable y_j, for each $j = 1, 2, ..., r$. There is no direct way of testing this assumption since the test would involve comparison of the observed values with the unobserved values for each dependent variable. Therefore, some caution must always be exercised in interpreting the results of the multivariate ratio analysis. Stanish, Gillings, and Koch [6] give an indirect method for testing the fundamental assumption and suggest a possible alternative strategy for analyzing the data if there are indications that the assumption is not true.

Finally, it should be emphasized that statistical manipulations are not a panacea for poor study management. Good management in clinical trials (and other experiments) should result in a minimal amount of missing data, in which case the methodology used in this paper is likely to be appropriate. When there is a substantial amount of missing data, say more than 10%, then the application of elaborate statistical strategies is probably not defensible. In this case, only a limited analysis should be undertaken for whatever data are available, and the results should be interpreted in a correspondingly limited context.

3. Use of MISCAT

Because MISCAT is an extension of GENCAT, and the use of the latter program has been thoroughly described elsewhere [5], this section is directed at details regarding the extended features of MISCAT and their relationship to previous features of GENCAT. In GENCAT terminology, the input mode of interest is raw data.

Suppose first there are no missing data. Means may be specified in the function definition stage in the following manner:

F(1) = MEAN(1)

F(2) = MEAN(2)

F(3) = MEAN(3) (3.1)

In this specification, the j-th function is defined, for $j = 1, 2, 3$, as the mean of the j-th dependent variable (on the input record), y_j. A maximum of 80 variables can be read in to define subpopulations and functions. Also, the program is currently written to accommodate a maximum of 80 functions. As in GENCAT, the information on each card begins in column one, and no blank spaces are allowed within the specification.

Strictly speaking, GENCAT has always computed means as the original functions of raw data input. However, in that setting, each function is a mean of scores based on a weighted or unweighted indicator function. If the indicator functions are unweighted, the resulting means are proportions. Specification of one or more cells of a conceptual multidimensional contingency table yields the proportions of subjects who are categorized to those cells. If the indicator functions are weighted by certain constants, the resulting means are simply the weighted proportions. Furthermore, if the response data are numerical (or ordinal), and the weights are equal to the respective observed values (or assigned scores) of some variable y_j, then the estimated marginal mean (or mean score) corresponding to y_j is obtained by summing these weighted proportions over the set of possible weights (observed values). Thus, the GENCAT user specifies contingency table cells and, optionally, weights as in the following example:

F(1) = G(1, 1)

F(2) = G(1, .) = W(1.0)

F(3) = G(2, .) = W(2.0)

F(4) = G(1, 1) + G(., 2) (3.2)

The $G(a_1, a_2)$ specification refers to the cell of the contingency table for which $y_j = a_j, j = 1, 2$. Thus, the first function is the proportion of subjects in the subpopulation under study for which $y_1 = 1$ and $y_2 = 1$. If a_j is replaced by a period in the specification, variable y_j is ignored. Thus, functions 2 and 3 are weighted proportions of subjects for which y_1 is equal to 1 and 2, respectively. If these are the only values y_1 can assume, then the estimated mean is obtained as F(2) + F(3). Finally, the symbol '+' is used to specify more than one cell of the contingency table. Thus, function 4 is the proportion of subjects for which ($y_1 = 1, y_2 = 1$) or $y_2 = 2$.

The MISCAT user can specify functions of the form (3.1) and (3.2). Moreover, MISCAT has greater flexibility than GENCAT, since the GENCAT specifications allow only 1 weight per function, whereas each mean specification of (3.1) assigns whatever weight is equal to the value of the observation.

If the MISCAT function definitions include contigency table cell specifications, the data for the dependent variables must be non-negative integers. However, if the only functions specified are means, as in (3.1), the data for the dependent variables may be any continuous data. In either case, the independent variables must still assume non-negative integer values in order to define the subpopulations under study. One important difference from GENCAT is that the end-of-data card supplied by the user must contain a negative integer in the first data field corresponding to an independent variable. In GENCAT, the negative integer was to be placed in the first data field, regardless of whether it corresponded to an independent or a dependent variable.

When there are missing data and the user specifies the missing data option, the multivariate ratio estimates, as described in section 2, are computed totally by MISCAT. The function definitions usually involve only mean specifications of the form (3.1), but they may also include contingency table cell specifications, as in (3.2). The methodology of section 2 is still well-defined in this case since these functions are also means. The functions which are averaged in the numerator of a ratio estimate are weighted indicator functions which assume the value of the weight if the observed data for a subject fall into one of the cells specified in a function definition. Otherwise, they assume the value zero. Note, however, in this multivariate case, that one or more variables might be unobserved, and yet the function value may still be nonzero. For example, function 3 in (3.2) may have the value of y_2 missing, but would still be nonzero if the value 2 were observed for y_1. The definitions of the missing data indicator functions (which are averaged to form the denominator of a ratio estimate) are not as straightforward in this multivariate case as they were in the univariate setting of section 2. The guiding principle is that the missing data indicator function assumes the value 1 (not missing) if it can be determined conclusively that the observed data for a subject either:

1. Fall into one of the cells specified for the function; or
2. Do not fall into any of the cells specified for the function.

Otherwise, it assumes the value zero (missing). Thus, for example, the value of function 1 in (3.2) is missing if the multivariate observation is $(M, 1)$, $(1, M)$, or (M, M) where M denotes a missing value. Otherwise, it could be determined whether or not the observation falls into the (1, 1) cell. The values of functions 2 and 3 in (3.2) are missing only if the value of y_1 is missing. The value of function 4 is missing if the multivariate observation is $(., M)$ or $(M, 1)$, where '.' denotes any value, observed or missing.

In order to specify the missing data option, the user simply adds one card between the function definition cards and the variable order card (see [5], p. 205). This card should have the words MISSING DATA in columns 1–12. Because an indicator function is created internally for each function specified by the user, the maximum number of functions which can be specified when using this option is 40 (or half as many as usual). Correspondingly, a maximum of 40 variables can be read in to define subpopulations and functions.

When the missing data option is specified, there is one additional input requirement. In the appropriate place on the raw data parameter card the user must specify an extended format by which the input data will be read. The extended format allows the data to be read twice, once in numeric mode, and once in character mode to detect missing values. As an illustration of how the extended format is writen, consider the following possible formats of input data:

(i) (F5.1, 10X, 3E8.2)

(ii) (8X, F5.4, 2(2X, E15.5)) (3.3)

The extended format expressions would be:

(i) (F5.1, 10X, 3E8.2, T1, A5, 10X, 3A8)

(ii) (8X, F5.4, 2(2X, E15.5), T1, 8X, A5, 2(2X, A15)) (3.4)

Note that the extended formats are constructed in the following manner:

1. The original format is extended first with specification T1 (tab to column 1), which resets the reader to column 1 of the input record.
2. All previous specifications are repeated, with the letters E and F replaced by the letter A, and with the deletion of each decimal point and each integer following a decimal point.

Since 'X' is a skip specification, the combination 'T1, 8X' is equivalent to 'T9', and thus, the second format in (3.4) may be shortened slightly.

It is important to note that if a data field is larger than 8 spaces (i.e., specified by Aw, where $w > 8$), then only the rightmost 8 spaces are checked to determine whether the field is blank. Thus, it is assumed that every non-missing data observation has a non-blank character somewhere within the rightmost 8 spaces of the data field. If this is not the case, then the extended format statement can be modified so as to read that part of the data field which always has non-blank characters for non-missing data. Thus, for example, 'A15' could be replaced by 'A8, 7X' if non-blank data appeared only in the first 8 spaces of the field.

Since the extended format is nearly twice as long as a usual format, the program has been revised so as to accommodate a continuation card, if necessary, following the raw data parameter card. The extended format may be continued in column one of the continuation card if it cannot fit entirely on the raw data parameter card.

Finally, when there are no missing data, it is possible to read several input records per subject by using the '/' specification in the format statement or by allowing the program to use the specified format repeatedly. However, when there are missing data, this option is not allowable, because the program expects to receive all of a subject's numeric values, followed by all of his character values. Thus, if the data set has several records per subject, it is necessary in the missing data case to create a new input data set with only one record per subject.

4. Example

As an illustration of the methodology and the use of MISCAT, an analysis is presented of data (see table 1) arising from a multi-center clinical trial for testing the efficacy and safety of a new drug for skin conditions. Patients were randomly assigned to drug or placebo in each of 6 clinics, and were evaluated prior to treatment to determine the initial severity of the skin condition. Finally, at 3 follow-up visits, patients were evaluated on a 5-point ordinal response scale representing extent of improvement. Thus, the response data lend themselves to analysis via multivariate mean scores.

Primary analyses of the data [6] show a significant treatment difference after adjustment for possible investigator effects. To illustrate the use of MISCAT, it is convenient to study response variability with respect to concomitant variables. For this purpose, the pooled sample of patients from all 6 clinics is used to analyze the data. The 3 time points are analyzed simultaneously in a multivariate setting in order to investigate the pattern of response over time. Similarly, it is of interest to examine the relationship between the initial stage of the disease and the subsequent evaluations of improvement. Because of the small number of patients whose initial disease stage was exacerbation, stages 4 and 5 were combined, reducing the number of initial disease stages under consideration to 2. These stages have been relabeled as moderate and severe. Finally, the quantitative responses associated with the response categories are the equal-increment scores shown in table 1.

The analysis involves 4 subpopulations, 2 treatments for each of 2 initial disease stages, and 3 means corresponding to the 3 follow-up visits. The 12 resulting functions in $\boldsymbol{R}$ and their estimated covariance matrix were computed by the methodology given in section 2. Shown in table 2 are the initial ratio estimates of the mean scores and corresponding standard errors. Analysis of $\boldsymbol{R}$ is undertaken with asymptotic regression models of the form (2.8).

A preliminary model of interest is the model which eliminates parameters corresponding to the effects of initial disease stage. This model is shown in table 3, together with parameter estimates and test results. The goodness of fit statistic indicates that the model is an adequate ($\alpha = 0.25$) representation of response variability. The design matrix $\boldsymbol{X}$ is constructed so as to have 1 module for each treatment. Within each of the 2 modules, indexed by $k = 1, 2$, there is 1 intercept (μ_k) and 1 parameter for each increment: from times 1 to 2 (α_k), and from times

Table 1
Data from clinical trial

INV	TRT	Stage	R1	R2	R3	INV	TRT	Stage	R1	R2	R3	INV	TRT	Stage	R1	R2	R3
5	1	3	3		3	6	2	4	2	2	2	10	1	3	1	1	1
5	1	3	3	2	3	6	2	3	3	3		10	1	3	1	1	1
5	1	4	3	2	2	6	2	4	4	4		10	1	3	3	3	3
5	1	3	2	2	1	6	2	4	4	3	3	10	1	3	1	1	1
5	1	3	3	2	2	6	2	4	5			10	1	3	2	2	2
5	1	4	2	1	3	6	2	3	1		1	10	1	3	2	2	1
5	1	4	1	1	1	6	2	3	4	2	4	10	2	3	3	3	3
5	1	4	1	1	1	6	2	4	5			10	2	3	4	4	4
5	1	5	5			6	2	5	4	5		10	2	3	1	1	1
5	1	3	1	1	1	6	2	4	4	4	3	10	2	3	2	2	
5	1	4	4	4	4	6	2	5	3	4	4	10	2	3	2	2	2
5	1	4	3	1	1	6	2	4	4	3	3	10	2	3	4	4	
5	1	4	1	1	1	8	1	4		4	4	10	2	3	1	1	2

Table 1 (continued)

INV	TRT	Stage	R1	R2	R3	INV	TRT	Stage	R1	R2	R3	INV	TRT	Stage	R1	R2	R3
5	1	4	3	3	3	8	1	4	3	2	1	10	2	3	2	3	3
5	1	4	1	1	1	8	1	5	1		1	10	2	3	4	3	3
5	1	3	1	1		8	1	4	1	1	1	10	2	3	3	3	3
5	1	3	4	4	4	8	1	3	2	1		10	2	4	3	3	4
5	1	3	3			8	1	4	2	1	1	10	2	3	3	3	4
5	1	4		1		8	1	3	1	1	1	10	2	3	3	3	3
5	2	3	4	3	3	8	1	4	2	2	2	10	2	3	5		
5	2	3	4	4	4	8	1	3	1	1	1	10	2	3	2	2	1
5	2	4	4	5	4	8	1	4	3	3	4	10	2	3	4	4	4
5	2	3	4	4	5	8	1	3	2	2	1	10	2	3	4	3	3
5	2	3	4	4	4	8	1	3	2	1	1	11	1	4	2	1	1
5	2	4	4	4	4	8	1	4	2	1	1	11	1	3	4	3	3
5	2	4	4			8	1	4	2	2	2	11	1	5	3		
5	2	3	4	4		8	1	4	3	2	1	11	1	3	2	1	1
5	2	3	2	2		8	1	4	2	1	1	11	1	4		3	2
5	2	5	3	3	4	8	1	4	2	2	1	11	1	4	3		
5	2	3	4	4	4	8	2	3	1	1	2	11	1	4	2	2	2
5	2	3	4	4		8	2	4	2	2	3	11	1	4	2	2	2
5	2	4	4	4		8	2	3	2	2	3	11	1	4	2	2	1
5	2	4	4	5		8	2	3	3	5	5	11	1	5	2	1	1
5	2	4	4	4		8	2	3	2	2	2	11	1	3	1	1	
5	2	3	4			8	2	4	3	3	3	11	1	3	2	1	1
5	2	4	1	1		8	2	3	3	3	3	11	1	3	3	2	2
5	2	3	4	4	4	8	2	5	4	3	3	11	1	5	2	2	1
6	1	3	3	3	3	8	2	4	4	4	5	11	1	5	1	1	1
6	1	4	2	2	2	8	2	5	4			11	1	4	2	1	1
6	1	4	3	2	2	8	2	3	3		5	11	2	4	2	2	1
6	1	4	4			8	2	5	4	3	4	11	2	4	4	4	4
6	1	4	2	2	2	8	2	3	2	3	3	11	2	4	4	4	4
6	1	4	2	2	1	9	1	5	2	2	1	11	2	4	4	3	4
6	1	4	3	3	3	9	2	4	3	3	3	11	2	3	4	4	
6	1	3	1	1	1	9	2	4	3	3	3	11	2	4	4	3	3
6	1	4	3	1	1	9	2	5	4	3	3	11	2	4	2	2	2
6	1	4	2	2	1	10	1	3	1	1	1	11	2	3	4	4	
6	1	3	2		1	10	1	3	1	1	1	11	2	5	4	3	3
6	1	3	3	4	4	10	1	3	2	2	1	11	2	4	4	3	3
6	1	5	2	2	2	10	1	3	2	2	1	11	2	4	3	3	3
6	1	4	2	1	1	10	1	3	1	1	1	11	2	4	2	2	1
6	1	4	3	4	4	10	1	3	3	2	1	11	2	3	4	3	3
6	1	4	1	1	1	10	1	3	2	2	2	11	2	4	4	4	4
6	1	4	1	1	1	10	1	3	1	1	1	11	2	3	4	4	3
6	2	4	3	3	3	10	1	3	3	1	1	11	2	4	4	3	3
6	2	4	4	4	4	10	1	3	2	2	2	11	2	3	4	3	3
6	2	4	2	2	2	10	1	3	3	2	2						
6	2	4	4	4		10	1	3	3	3	2						

INV, Investigator identification number (5, 6, 8, 9, 10, 11); TRT, Treatment (1 = test drug, 2 = placebo); Stage, Initial stage of disease (3 = fair, 4 = poor, 5 = exacerbation); R1, Response at Time 1; R2, Response at Time 2; R3, Response at Time 3; 1 = Rapidly Improving, 2 = Slowly Improving, 3 = Stable, 4 = Slowly Worsening, 5 = Rapidly Worsening, Blank = Missing Data.

Table 2
Observed and predicted mean scores for overall-evaluation and the corresponding estimated standard errors

Treatment	Initial disease stage	Time	Preliminary ratio estimates		Multivariate ratio analysis		Modified ratio analysis procedure	
			Mean score	Estimated standard error	Predicted mean score	Estimated standard error	Predicted mean score	Estimated standard error
Drug	Moderate	1	2.075	0.1431	2.151	0.0957	2.152	0.0971
Drug	Moderate	2	1.730	0.1412	1.784	0.0943	1.814	0.1019
Drug	Moderate	3	1.611	0.1484	1.629	0.0989	1.690	1.1060
Drug	Severe	1	2.222	0.1329	2.151	0.0957	2.152	0.0971
Drug	Severe	2	1.791	0.1338	1.784	0.0943	1.814	0.1019
Drug	Severe	3	1.651	0.1467	1.629	0.0989	1.690	0.1060
Placebo	Moderate	1	3.146	0.1674	3.347	0.0984	3.391	0.0993
Placebo	Moderate	2	3.054	0.1619	3.192	0.0974	3.267	0.1000
Placebo	Moderate	3	3.129	0.1922	3.192	0.0974	3.267	0.1000
Placebo	Severe	1	3.512	0.1331	3.347	0.0984	3.391	0.0993
Placebo	Severe	2	3.256	0.1438	3.192	0.0974	3.267	0.1000
Placebo	Severe	3	3.188	0.1559	3.192	0.0974	3.267	0.1000

2 to 3 (β_k). The test results suggest that 2 of these increments (β_1, α_2) are equal while a third (β_2) is zero ($Q = 0.42, d.f. = 2$), and hence the final model incorporates these constraints into the design matrix.

Shown in table 3 is the final model, together with parameter estimates and test results which indicate the following conclusions with respect to the response variable, extent of improvement.

1. The goodness of fit test ($Q = 4.83, d.f. = 8$) suggests that the model is an adequate representation of the data.

2. There are 2 distinct levels of improvement at the first time point, 1 for patients on drug and 1 for patients on placebo. These 2 levels are represented by the first and last columns of the design matrix. The result $\mu_1 \neq \mu_2$ indicates a level of improvement significantly better for patients on drug.

3. Drug is significantly better than placebo at each time as well as for all 3 time points considered jointly.

4. During the time intervals subsequent to time one, there are three distinct levels of the extent of additional improvement, as follows:

Treatment group	Time interval	Extent of additional improvement
Drug	1–2	α_1
Drug	2–3	β_1
Placebo	1–2	β_1
Placebo	2–3	0

The result $\alpha_1 \neq \beta_1$ indicates 2 conclusions:

(i) Patients on drug improve more during time interval 1–2 than during 2–3. This is not surprising, since at time 2, many patients may be nearly cured, in which case there would be little room for further improvement.

(ii) Patients on drug improve more in time interval 1–2 than placebo patients during any time interval.

It should be remembered that the improvement scores are subjective ratings by the investigators, and thus, the conclusions are conditional on the reliability and validity of the data. No analysis of these issues is undertaken here.

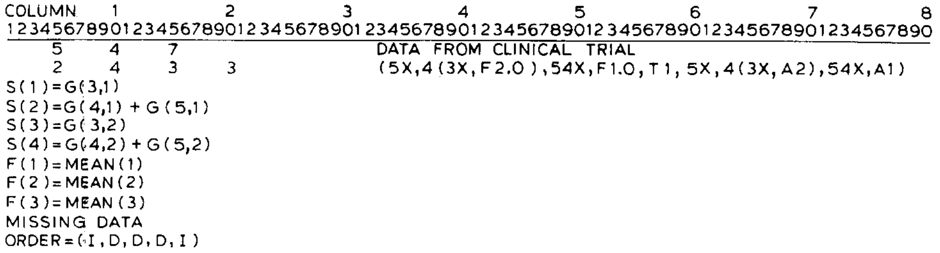
Shown in table 2 under the heading, Multivariate ratio analysis, are the predicted mean scores based on the final model, together with their estimated standard errors which are uniformly smaller than the corresponding preliminary estimates. In addition, it is clear that the mean scores predicted from the

Table 3
Summary of results from multivariate ratio analysis

Preliminary model	Parameters	Estimates	Standard errors	Hypotheses	d.f.	Q statistics
$X = \left[\begin{array}{cccccc} 1&0&0&&& \\ 1&1&0&&& \\ 1&1&1&&& \\ 1&0&0&&& \\ 1&1&0&&& \\ 1&1&1&&& \\ &&&1&0&0 \\ &&&1&1&0 \\ &&&1&1&1 \\ &&&1&0&0 \\ &&&1&1&0 \\ &&&1&1&1 \end{array}\right]$	μ_1	2.150	0.096	$\beta_1 = \beta_2 = 0$	2	4.55
	α_1	−0.373	0.078	$\alpha_1 = \alpha_2 = 0$	2	29.05 [c]
	β_1	−0.130	0.061	$\alpha_1 = \alpha_2, \beta_1 = \beta_2$	2	5.36 [a]
	μ_2	3.366	0.104	$\alpha_1 = \beta_1, \alpha_2 = \beta_2$	2	6.88 [b]
	α_2	−0.195	0.080	$\beta_1 = \alpha_2, \beta_2 = 0$	2	0.42
	β_2	0.008	0.101	No lack of fit	6	4.41

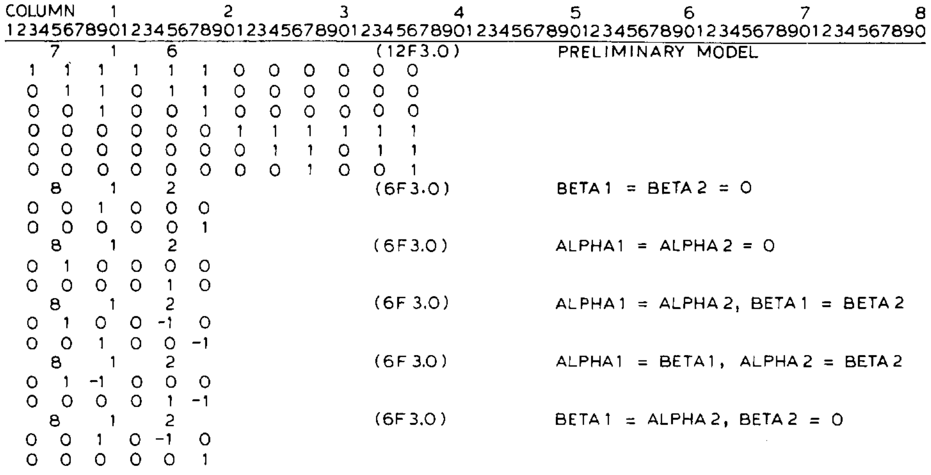
Final model	Parameters	Estimates	Standard errors	Hypotheses	d.f.	Q statistics
$X = \left[\begin{array}{cccc} 1&0&0& \\ 1&1&0& \\ 1&1&1& \\ 1&0&0& \\ 1&1&0& \\ 1&1&1& \\ &&0&1 \\ &&1&1 \\ &&1&1 \\ &&0&1 \\ &&1&1 \\ &&1&1 \end{array}\right]$	μ_1	2.151	0.096	Equal increments ($\alpha_1 = \beta_1$)	1	4.77 [b]
	α_1	−0.367	0.077	No treatment effect time 1 ($\mu_1 = \mu_2$)	1	76.31 [c]
	β_1	−0.155	0.048	No treatment effect time 2 ($\mu_1 + \alpha_1 = \mu_2 + \beta_1$)	1	104.16 [c]
	μ_2	3.347	0.098	No treatment effect time 3 ($\mu_1 + \alpha_1 = \mu_2$)	1	136.87 [c]
				No treatment effect (also total variation) ($\mu_1 = \mu_2, \alpha_1 = \beta_1 = 0$)	3	148.03 [c]
				No lack of fit	8	4.83

[a] Denotes significance at $\alpha = 0.10$.
[b] Denotes significance at $\alpha = 0.05$.
[c] Denotes significance at $\alpha = 0.01$.



```
COLUMN   1         2         3         4         5         6         7         8
12345678901234567890123456789012345678901234567890123456789012345678901234567890
    5    4    7                 DATA FROM CLINICAL TRIAL
    2    4    3    3            (5X,4(3X,F2.0),54X,F1.0,T1,5X,4(3X,A2),54X,A1)
S(1)=G(3,1)
S(2)=G(4,1)+G(5,1)
S(3)=G(3,2)
S(4)=G(4,2)+G(5,2)
F(1)=MEAN(1)
F(2)=MEAN(2)
F(3)=MEAN(3)
MISSING DATA
ORDER=(I,D,D,D,I)
```

Fig. 1.



```
COLUMN   1         2         3         4         5         6         7         8
12345678901234567890123456789012345678901234567890123456789012345678901234567890
  7  1  6                        (12F3.0)           PRELIMINARY MODEL
  1  1  1  1  1  1  0  0  0  0  0  0
  0  1  1  0  1  1  0  0  0  0  0  0
  0  0  1  0  0  1  0  0  0  0  0  0
  0  0  0  0  0  0  1  1  1  1  1  1
  0  0  0  0  0  0  0  1  1  0  1  1
  0  0  0  0  0  0  0  0  1  0  0  1
  8  1  2                         (6F3.0)           BETA1 = BETA2 = 0
  0  0  1  0  0  0
  0  0  0  0  0  1
  8  1  2                         (6F3.0)           ALPHA1 = ALPHA2 = 0
  0  1  0  0  0  0
  0  0  0  0  1  0
  8  1  2                         (6F3.0)           ALPHA1 = ALPHA2, BETA1 = BETA2
  0  1  0  0 -1  0
  0  0  1  0  0 -1
  8  1  2                         (6F3.0)           ALPHA1 = BETA1, ALPHA2 = BETA2
  0  1 -1  0  0  0
  0  0  0  0  1 -1
  8  1  2                         (6F3.0)           BETA1 = ALPHA2, BETA2 = 0
  0  0  1  0 -1  0
  0  0  0  0  0  1
```

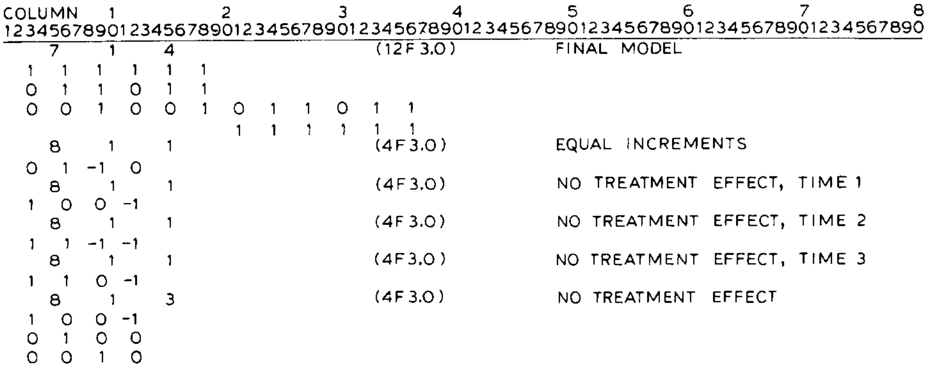
Fig. 2.

multivariate ratio analysis provide a good fit to the original ratio estimates.

The parameter cards for MISCAT which produce the estimates and test results given in the tables are shown in fig. 1–3, and are described in the following paragraphs. Refer to the GENCAT paper [5] for detailed explanations of the parameter cards. The basic parameter card (see fig. 1) indicates that the type of input is raw data, which is read from unit 7 (disk in this case). The next card specifies the extended format of the input data, the number of subpopulations (4) and functions (3) to be formed, and the number of dependent (3) and independent (2) variables to be input.

These cards are followed by subpopulation definition cards, which indicate how the 4 subpopulations are to be formed on the basis of the 2 independent variables, and function definition cards, which are the mean specifications corresponding to (3.1). The missing data card is next, and the last parameter card describing the input data is a card which specifies the order of the dependent and independent variables on the input record.

The first card necessary to fit the preliminary asymptotic regression model to the ratio estimates is a design matrix parameter card (see fig. 2) indicating the number (6) of columns in the X matrix of (2.8). This is followed by a set of 6 cards corresponding to



```
COLUMN   1         2         3         4         5         6         7         8
12345678901234567890123456789012345678901234567890123456789012345678901234567890
  7  1  4                        (12F3.0)           FINAL MODEL
  1  1  1  1  1  1
  0  1  1  0  1  1
  0  0  1  0  0  1  0  1  1  0  1  1
                    1  1  1  1  1  1
  8  1  1                         (4F3.0)           EQUAL INCREMENTS
  0  1 -1  0
  8  1  1                         (4F3.0)           NO TREATMENT EFFECT, TIME 1
  1  0  0 -1
  8  1  1                         (4F3.0)           NO TREATMENT EFFECT, TIME 2
  1  1 -1 -1
  8  1  1                         (4F3.0)           NO TREATMENT EFFECT, TIME 3
  1  1  0 -1
  8  1  3                         (4F3.0)           NO TREATMENT EFFECT
  1  0  0 -1
  0  1  0  0
  0  0  1  0
```

Fig. 3.

the columns of the matrix, and containing the matrix values. Finally, there are sets of contrast matrix cards which test the hypotheses in table 3 corresponding to the preliminary model. The first card of each set specifies the number of rows of the matrix $\boldsymbol{C}$, where the hypothesis is $\boldsymbol{C\beta} = \boldsymbol{0}$, and $\boldsymbol{\beta}$ is the vector of model parameters. The last cards of each set contain the values of the $\boldsymbol{C}$-matrix.

The cards shown in fig. 3 specify the final model and the corresponding hypothesis tests. Their format and arrangement are similar to that of the cards shown in fig. 2. These cards may follow those of fig. 2 in the same computer run, or they may replace the cards of fig. 2 in a subsequent run.

5. Hardware specifications

MISCAT is written in double precision in IBM System 360/370 FORTRAN IV which incorporates a few extensions to American National Standard (ANS) FORTRAN. As a result, minor modifications of the source code may be required to use the program on other machines.

6. Program availability

The source deck for MISCAT, together with the corresponding listing, may be obtained for a nominal cost from the Department of Biostatistics, School of Public Health, University of Michigan, Ann Arbor, MI 48109, USA. The current documentation and operating instructions are included with the initial purchase of the program. Furthermore, purchasers may place their names on an active mailing list to receive information concerning updating and further modifications of the program.

7. Disclaimer

Although MISCAT has been tested extensively, no warranty, expressed or implied, is made to the accuracy and functioning of the program. No responsibility is assumed by the authors. However, if specific problems or questions do arise, contact Dr J. Richard Landis at the Department of Biostatistics, School of Public Health, University of Michigan, Ann Arbor, MI 48109, USA.

Acknowledgments

The authors would like to thank Hoechst-Roussel Pharmaceuticals, Inc. for providing the data which were used for the example in this paper. The authors also thank Jean McKinney, Jean Harrison, and Cristina Ramirez for their conscientious typing of the original manuscript and subsequent revisions.

References

[1] W.G. Cochran, Sampling Techniques, 2nd edn (John Wiley and Sons, New York, 1963).
[2] J. Cornfield, J. Am. Stat. Assoc. 39 (1944) 236–239.
[3] G.G. Koch, D.H. Freeman and J.L. Freeman, Int. Stat. Rev. 43 (1975) 59–78.
[4] G.G. Koch, J.R. Landis, J.L. Freeman, D.H. Freeman and R.G. Lehnen, Biometrics 33 (1977) 133–158.
[5] J.R. Landis, W.M. Stanish, J.L. Freeman and G.G. Koch, Comput. Prog. Biomed. 6 (1976) 196–231.
[6] W.M. Stanish, D.B. Gillings, and G.G. Koch, Biometrics 34 (1978) in press.
[7] A. Wald, Trans. Am. Math. Soc. 54 (1943) 426–482.