Computer Programs in Biomedicine 6 (1976) 196–231
© North-Holland Publishing Company

A COMPUTER PROGRAM FOR THE GENERALIZED CHI-SQUARE ANALYSIS OF CATEGORICAL DATA USING WEIGHTED LEAST SQUARES (GENCAT)

J. Richard LANDIS [a], William M. STANISH [b], Jean L. FREEMAN [c] and Gary G. KOCH [b]

[a] *Dept. of Biostatistics, University of Michigan, Ann Arbor, Michigan 48109;* [b] *Dept. of Biostatistics, University of North Carolina, Chapel Hill, North Carolina 27514;* [c] *Yale University School of Medicine, New Haven, Connecticut 06510, USA*

GENCAT is a computer program which implements an extremely general methodology for the analysis of multivariate categorical data. This approach essentially involves the construction of test statistics for hypotheses involving functions of the observed proportions which are directed at the relationships under investigation and the estimation of corresponding model parameters via weighted least squares computations. Any compounded function of the observed proportions which can be formulated as a sequence of the following transformations of the data vector – linear, logarithmic, exponential, or the addition of a vector of constants – can be analyzed within this general framework. This algorithm produces minimum modified chi-square statistics which are obtained by partitioning the sums of squares as in ANOVA. The input data can be either: (a) frequencies from a multidimensional contingency table; (b) a vector of functions with its estimated covariance matrix; and (c) raw data in the form of integer-valued variables associated with each subject. The input format is completely flexible for the data as well as for the matrices.

Multivariate analysis Categorical data Contingency tables Minimum modified chi-square Weighted least squares Linear models Computer program Rates and proportions

1. Introduction

The analysis of multivariate categorical data has received considerable attention in recent years. In particular, the methodology proposed in Grizzle, Starmer, and Koch [1] (hereafter abbreviated GSK) has been extended to provide models for a wide variety of statistical problems as discussed in [2–20]. This approach essentially involves a two stage procedure:

(i) the construction of the appropriate functions of the observed proportions which are directed at the relationships under investigation by a sequence of matrix operations, together with logarithmic and exponential transformations;

(ii) the construction of test statistics for hypotheses involving these functions and the estimation of corresponding model parameters via weighted least squares computations.

The basic elements of the theoretical justification for this methodology are given in Section 2.

In principle, any compounded function of the observed proportions associated with the multidimensional contingency table which can be formulated by successive transformations – either linear, logarithmic, exponential, or addition of a vector of constants – of the data vector can be analyzed within this framework. On the other hand, all the models considered in the original GSK paper [1] could be expressed as either linear or log-linear functions of the observed proportions. As a result, the corresponding computer programs CATLIN and LINCAT discussed in [21] were developed for analyzing functions limited to the scope encompassed by these two classes.

Subsequent to these developments, Forthofer and Koch [7] extended the GSK procedure to include two more general classes of compounded functions of the observed proportions which permitted the investigation of more complex relationships in the data. Accordingly, they provided a computer program MODCAT discussed in [22] which can be used to implement models involving compounded functions of the types specified in [7].

Even though these classes of functions specified in [1,7] are quite adequate for a wide range of statistical problems, there are a number of situations in which the functions of interest cannot be expressed as one of these standard types. For example, see [8,17,18]. Moreover, in many large data sets the size of the underlying contingency table is outside the scope of computational feasibility for the GSK approach to be applied directly. In such situations, specialized computing procedures are required to obtain the estimates of the pertinent functions and their estimated covariance matrix from the raw data associated with each subject. As discussed in [9,20] the same estimators which would need to be obtained from the conceptual multidimensional contingency table can be generated by computing the across-subject arithmetic means of appropriately chosen indicator functions. A summary of these indicator function techniques is given in Section 3.

Consequently, in order to implement the analyses of an extremely general class of compounded functions and to provide more flexible options for data input, a new computer program GENCAT has been developed. In addition to permitting the analysis of more general functions, this program can handle input data from either cards, tape, or disk file in any of the following three different forms:

(i) observed frequencies from a multidimensional contingency table;

(ii) a vector of functions with its estimated covariance matrix;

(iii) raw data in the form of non-negative integer-valued variables associated with each subject.

The format for the data and the linear operator matrices can be specified separately for each set of data and for each matrix as alternatives to the pre-specified default formats.

Several analyses of a given data set can be performed in the same computer run by fitting more than one design matrix to a particular set of functions and by testing several contrast matrices for each model. In addition, the user may specify that a vector of functions and its estimated covariance matrix resulting from any one (and only one) of the following steps:

(a) the original data vector;

(b) the functions resulting from a particular transformation;

(c) the estimated model parameters obtained from a particular design matrix;

(d) the estimated parameters obtained from a particular contrast matrix, be saved for reanalysis at a later stage in the same run or be written to a file to be used in a subsequent computer run. Finally, multiple sets of data can be processed in the same computer run by simply repeating the appropriate sequence of cards described in section 5.

2. Statistical theory

Let $j = 1, 2, ..., r$ index a set of categories which correspond to the r response profiles associated with the specific dependent variables of interest. For example, in a multidimensional contingency table with two binary dependent variables (Y_1, Y_2), the $r = 4$ response profiles are (0,0), (0,1), (1,0), (1,1). Similarly, let $i = 1, 2, ..., s$ index a set of categories which correspond to distinct sub-populations as defined in terms of pertinent independent variables. If samples of size n_i where $i = 1, 2, ..., s$ are independently selected from the respective sub-populations, then the resulting data can be summarized in an $(s \times r)$ contingency table as shown in table 1, where n_{ij} denotes the frequency of response category j in the sample from the i-th sub-population.

Table 1
Observed contingency table

Sub-population	Response profile categories				
	1	2	...	r	Total
1	n_{11}	n_{12}	...	n_{1r}	n_1
2	n_{21}	n_{22}	...	n_{2r}	n_2
...	...	...	...	...	...
s	n_{s1}	n_{s2}	...	n_{sr}	n_s

The vector $\boldsymbol{n}_i$ where $\boldsymbol{n}_i' = (n_{i1}, n_{i2}, ..., n_{ir})$ will be assumed to follow the multinomial distribution with parameters n_i and $\boldsymbol{\pi}_i' = (\pi_{i1}, \pi_{i2}, ..., \pi_{ir})$, where π_{ij} represents the probability that a randomly selected element from the i-th sub-population is classified in the j-th response category. Thus, the relevant product multinomial model is

$$\phi = \prod_{i=1}^{s} \left\{ n_i! \prod_{j=1}^{r} [\pi_{ij}^{n_{ij}}/n_{ij}!] \right\} \qquad (2.1)$$

with the constraints

$$\sum_{j=1}^{r} \pi_{ij} = 1 \text{ for } i = 1, 2, ..., s. \qquad (2.2)$$

Let $p_i = (n_i/n_i)$ be the $(r \times 1)$ vector of observed proportions associated with the sample from the i-th sub-population and let p be the $(sr \times 1)$ compound vector defined by $p' = (p'_1, p'_2, ..., p'_s)$. Thus, the vector p is the unrestricted maximum likelihood estimator of π where $\pi' = (\pi'_1, \pi'_2, ..., \pi'_s)$. A consistent estimator for the covariance matrix of p is given by the $(sr \times sr)$ block diagonal matrix $V(p)$ with the matrices

$$\underset{(r \times r)}{V_i (p_i)} = \frac{1}{n_i} [D_{p_i} - p_i p'_i] \tag{2.3}$$

for $i = 1,2, ..., s$ on the main diagonal, where D_{p_i} is an $(r \times r)$ diagonal matrix with elements of the vector p_i on the main diagonal.

Let $F_1(p), F_2(p), ..., F_u(p)$ be a set of u functions of p which pertain to some aspect of the relationship between the distribution of the response profiles and the nature of the sub-populations. Each of these functions is assumed to have continuous partial derivatives through order two with respect to the elements of p within an open region containing $\pi = E\{p\}$. If $F \equiv F(p)$ is defined by

$$F' = [F(p)]' = [F_1(p), F_2(p), ..., F_u(p)], \tag{2.4}$$

then a consistent estimator for the covariance matrix of F is the $(u \times u)$ matrix

$$V_F = H[V(p)]H', \tag{2.5}$$

where $H = [dF(x)/dx | x = p]$ is the $(u \times sr)$ matrix of first partial derivatives of the functions F evaluated at p. In all applications, the functions comprising F are chosen so that V_F is asymptotically nonsingular.

The function vector F is a consistent estimator of $F(\pi)$. Hence, the variation among the elements of $F(\pi)$ can be investigated by fitting linear regression models by the method of weighted least squares. This phase of the analysis can be characterized by writing

$$E_A\{F\} \equiv E_A\{F(p)\} = F(\pi) = X\beta, \tag{2.6}$$

where X is a pre-specified $(u \times t)$ design (or independent variable) matrix of known coefficients with full rank $t \leqslant u$, β is an unknown $(t \times 1)$ vector of parameters, and "E_A" means "asymptotic expectation."

As discussed in more detail in Koch et al. [9], an appropriate test statistic for the goodness of fit of the model (2.6) is

$$Q = Q(X, F) = (R\,F)'\,[R\,V_F R']^{-1}\,R\,F, \tag{2.7}$$

where R is any full rank $[(u - t) \times u]$ matrix orthogonal to X. Here Q is approximately distributed according to the χ^2 distribution with $D.F. = (u - t)$, if the sample sizes $\{n_i\}$ are sufficiently large that the elements of the vector F have an approximate multivariate normal distribution as a consequence of Central Limit Theory (CLT). Test statistics such as Q are known as generalized Wald [23] statistics and various aspects of their application to a broad range of problems involving the analysis of multivariate categorical data are discussed in Bhapkar and Koch [24,25] and Grizzle et al. [1].

However, these test statistics like (2.7) are obtained in actual practice by using weighted least squares as a computational algorithm which is justified on the basis of the fact that Q of (2.7) is identically equal to

$$Q = (F - X\,b)'\,V_F^{-1}\,(F - X\,b), \tag{2.8}$$

where

$$b = (X'\,V_F^{-1}\,X)^{-1}\,X'\,V_F^{-1}\,F \tag{2.9}$$

is a BAN estimator for β based on the linearized modified χ_1^2-statistic of Neyman [26]. In view of this identity demonstrated in Bhapkar [27], both Q and b are regarded as having reasonable statistical properties in samples which are sufficiently large for applying CLT to the functions F. As a result, a consistent estimator for the covariance matrix of b is given by

$$V_b = (X'\,V_F^{-1}\,X)^{-1}. \tag{2.10}$$

If the model (2.6) does adequately characterize the vector $F(\pi)$, tests of linear hypotheses pertaining to the parameters β can be undertaken by standard multiple regression procedures. In particular, for a general hypothesis of the form

$$H_0: C\beta = O, \tag{2.11}$$

where C is a known $(c \times t)$ matrix of full rank $c \leqslant t$ and O is a $(c \times 1)$ vector of O's, a suitable test statistic is

$$Q_C = (C\,b)'\,[C\,(X'\,V_F^{-1}\,X)^{-1}\,C']^{-1}\,C\,b \qquad (2.12)$$

which has approximately a χ^2-distribution with *D.F.* $= c$ in large samples under H_0 in (2.11).

In this framework, the test statistic Q_C reflects the amount by which the goodness of fit statistic (2.8) would increase if the model (2.6) were simplified (or reduced) by substitutions based on the additional constraints implied by (2.11). Thus, these methods permit the total variation within $F(\pi)$ to be partitioned into specific sources and hence represent a statistically valid analysis of variance for the corresponding estimator functions F.

Predicted values for $F(\pi)$ based on the model (2.6) can be calculated from

$$\hat{F} = X\,b = X\,(X'\,V_F^{-1}\,X)^{-1}\,X'\,V_F^{-1}\,F. \qquad (2.13)$$

Thus, consistent estimators for the variances of the elements of $\hat{F}$ can be obtained from the diagonal elements of

$$V_{\hat{F}} = X\,(X'\,V_F^{-1}\,X)^{-1}\,X'. \qquad (2.14)$$

The predicted values $\hat{F}$ not only have the advantage of characterizing essentially all the important features of the variation in $F(\pi)$, but also represent better estimators than the original function statistics F since they are based on the data from the entire sample as opposed to its component parts. Moreover, they are descriptively advantageous in the sense that they make trends more apparent and permit a clearer interpretation of the relationship between $F(\pi)$ and the variables comprising the columns of X.

Although the formulation of $F(p)$ can be quite general, Grizzle et al. [1] and Forthofer and Koch [7] demonstrated that a wide range of problems in categorical data analysis could be considered within the framework of a few specified classes of compounded logarithmic, exponential, and linear functions. However, these functions are all special cases of a broad class of functions which can be expressed in terms of repeated applications of any sequence of the following matrix operations:

(i) linear transformation of the type

$$F_1(p) = A_1\,p = a_1, \qquad (2.15)$$

where A_1 is a matrix of known constants;

(ii) logarithmic transformations of the type

$$F_2(p) = \boldsymbol{log}_e(p) = a_2, \qquad (2.16)$$

where $\boldsymbol{log}_e$ transforms a vector to the corresponding vector of natural logarithms;

(iii) exponential transformations of the type

$$F_3(p) = \boldsymbol{exp}(p) = a_3, \qquad (2.17)$$

where $\boldsymbol{exp}$ transforms a vector to the corresponding vector of exponential functions, i.e., of anti-logarithms. Then the linearized Taylor-series-based estimate of the covariance matrix of F_k for $k = 1, 2, 3$, is given by (2.5), where the corresponding H_k matrix operator is

$$H_1 = A_1; \qquad (2.18)$$

$$H_2 = D_p^{-1}; \qquad (2.19)$$

$$H_3 = D_{a_3}, \qquad (2.20)$$

where D_y is a diagonal matrix with elements of the vector y on the main diagonal. As a result, an extremely general class of functions of the observed proportions can be formulated by successively compounding transformations of the types in (2.15)–(2.17) in any desired order. Moreover, the linearized Taylor-series-based estimate of the covariance matrix associated with a given set of compounded functions can be obtained by repeated application of the chain rule for matrix differentiation. Examples of the types of compounded functions which are useful in specific statistical applications are discussed in Section 4.

3. Data input

The data input options for GENCAT are quite flexible, permitting the data set to be read in from any input device such as cards, tape, or disk files. If the data are already in the form of a contingency table having r response profiles within each of s sub-populations as shown in table 1, the observed frequencies can be handled in one of two slightly different ways, depending on the size of $r * s$. (In either case, though, the requirement is $r \leqslant 80, p \leqslant 80$).

3.1. Frequency data: CASE 1

If $r * s \leqslant 80$, the observed frequencies in the format of table 1 are entered in row order. As a result, the data vector $\boldsymbol{p}$ is of dimension $r * s$ and consists of the r proportions within each of the s sub-populations. This compound vector $\boldsymbol{p}$ can then be analyzed directly with the usual constraint that the final set of functions $F(\boldsymbol{p})$ must have a non-singular covariance matrix V_F. This input mode is similar to that of the previous programs CATLIN [21] and MODCAT [22].

3.2. Frequency data: CASE 2

If $r * s > 80$, the observed frequencies in the format of table 1 are also entered in row order with the condition that the same linear functions are to be formed within each sub-population at the initial step so that the total number of functions does not exceed 80. As a result, the data vector $\boldsymbol{p}$ is of dimension $r' * s \leqslant 80$, where r' is the number of linear combinations of the proportions associated with the r response profiles used to form the reduced number of functions within each sub-population. This compound vector $\boldsymbol{p}$ can then be analyzed directly with the usual constraint that the final set of functions $F(\boldsymbol{p})$ must have a non-singular covariance matrix V_F. This input mode is similar to that of the previous program Lincat [21] which required a block diagonal A matrix at the initial stage of function formulation.

If $r * s > 80$ and the first set of functions of the underlying proportions cannot be obtained by using a block diagonal linear operator matrix with the same block for each sub-population, then the current fixed dimension of the program must be increased to accommodate the larger number of functions. Changing the dimensions of the program is not difficult. Instructions are included in the distribution package (see Section 9).

3.3. Direct input of function vector

In some situations the observed frequencies cannot be assumed to follow the multinomial distribution specified in (2.1). Consequently, the estimate of the covariance matrix in (2.3) may not be appropriate. For example, the analysis of data from sample surveys as discussed in [11,13] requires estimates of the covariance matrix which are consistent with more complex sampling schemes and may need to be computed via alternative procedures such as balanced repeated replication. Also, for modularized analyses involving preliminary estimates of functions within each module as discussed in [15,16] the covariance structure of the functions may be estimated by iterative procedures associated with maximum likelihood estimation. For such cases, the input data are already in the form of a vector of functions and its estimated covariance matrix. Accordingly, these data can be entered directly for the vector $F(\boldsymbol{p})$ and its corresponding covariance matrix V_F. Thus, subsequent analyses of these functions can be performed directly.

3.4. Raw data

If the size of the underlying multidimensional contingency table in the format of table 1 is outside the scope of computational feasibility for a specific problem, the data can be analyzed in an alternative mode which effectively bypasses the construction of a contingency table. In such instances the raw data associated with each subject or observational unit can be entered in the form of categorical variables which are classified as either independent or dependent. The independent variables are used to specify the s sub-populations, whereas the dependent variables are used to form the u response functions within each sub-population. Weights are assigned to each of u indicator variables so that the across-subject arithmetic means of these variables provide the required estimators. In particular, let the response functions selected for analysis be indexed by $k = 1, 2, ..., u$. These functions are expressed in terms of combinations of specified levels of the dependent variables. For this purpose, let the u indicator variables be defined by

$$z_{ikl} = \begin{cases} w_{k(j)}, & \text{if the } l\text{-th subject in the sample from the } i\text{-th sub-population is classified into response profile } j \\ O, & \text{otherwise,} \end{cases} \qquad (3.1)$$

for $i = 1, 2, ..., s; j = 1, 2, ..., r; k = 1, 2, ..., u; l = 1, 2, ..., n_i$, where $w_{k(j)}$ is the weight assigned to the j-th response profile for the k-th function. Thus, the mean scores for the u functions within the s sub-populations can be obtained directly from (3.1) as

$$\bar{z}_{ik} = \frac{1}{n_i} \sum_{l=1}^{n_i} z_{ikl}. \tag{3.2}$$

For notational convenience, these indicator functions can be summarized in vector notation by letting

$$z'_{il} = (z_{i1l}, z_{i2l}, ..., z_{iul}); \tag{3.3}$$

$$\bar{z}'_i = (\bar{z}_{i1}, \bar{z}_{i2}, ..., \bar{z}_{iu}); \tag{3.4}$$

$$\bar{z}' = (\bar{z}'_1, \bar{z}'_2, ..., \bar{z}'_s). \tag{3.5}$$

Thus, a consistent estimate of the covariance matrix of $\bar{z}$ in (3.5) can be obtained from (3.3) and (3.4) by computing

$$V(\bar{z}_i) = \frac{1}{n_i^2} \sum_{l=1}^{n_i} (z_{il} - \bar{z}_i)(z_{il} - \bar{z}_i)' \tag{3.6}$$

for $i = 1, 2, ..., s$, and then constructing a block diagonal matrix $V(\bar{z})$ with the matrices in (3.6) as the corresponding blocks on the main diagonal.

In this computer program all the variables must be categorical with non-negative integer-valued codes, e.g., 0, 1, 2, ..., L. If the data are not already expressed in this form, a general purpose statistical package such as SPSS [28] can be used to recode the data prior to using GENCAT. When using the raw data input mode, all the computations in (3.2) and (3.6) are automatically performed for the specified indicator functions in (3.1). Then the vector $\bar{z}$ in (3.5) and its estimated covariance matrix $V(\bar{z})$ obtained from (3.6) are used for F and V_F in the original GSK framework. Specific applications of these alternative techniques involving indicator variables are discussed further in [9,18,20].

4. Formulation of functions

Since the GSK approach to the analysis of multivariate categorical data is extremely general, the user must specify the set of functions aimed at the relationships under investigation. Depending on the type of input data, these functions are derived either from the vector of observed proportions associated with contingency table frequencies, the vector of direct input functions, or from the raw data associated with each subject. These functions are formed by successive applications of linear operator matrices $A_1, A_2, ..., A_{m_1}$ and transformations $T_1, T_2, ..., T_{m_2}$, where T_m transforms each element of a vector via logarithms or exponentiation as shown in (2.16) and (2.17), or by the addition of a vector of constants. In general, these compounded functions of the observed proportions can be expressed as

$$F(p) = T_{m_2}(A_{m_1} * ... * T_2(A_2 * T_1(A_1 * p)) ...), \tag{4.1}$$

where p is the vector of proportions created at the initial stage of data input.

The estimation of these compounded functions can sometimes lead to computational difficulties. For example, the logarithmic transformation can be applied only to function vectors in which all the elements are positive. Furthermore, the linear operator matrices and the transformations must be chosen in such a way that the estimated covariance matrix associated with the final set of functions in non-singular. In this regard, the presence of zero cells in the underlying multidimensional contingency table may cause a problem for some applications. For these cases, such zero frequencies can occasionally be replaced by (1/2) in the computations as proposed in [1] without adversely affecting the validity of the analysis, provided that the total number of such adjustments is relatively small. Otherwise, some linear functions such as mean scores or first-order marginal sums are obtained by combining observed proportions across categories, and thus individual zero cells are not necessarily troublesome unless they induce functional dependence among certain elements of $F(p)$. Finally, a more detailed discussion of the impact of zero cells is given in [9,17].

Although $F(p)$ in (4.1) includes a wide range of possible functions, many of the relationships commonly investigated for multivariate categorical data involve functions which can be expressed in one of the following classes.

4.1. Linear functions

In many situations, the hypotheses of interest are expressed in terms of constraints on the observed cell probabilities, on the first or higher-order marginal distributions, or on mean scores associated with the response variables. For example, most of the relevant hypotheses involving marginal distributions and/or mean scores in repeated measurement designs, hypotheses involving main effects and interactions in factorial designs, and hypotheses involving incomplete data and supplemented margins can all be expressed in terms of linear functions of the observed proportions as

$$F(p) = A_1 * p. \tag{4.2}$$

For further details, see [1–3,5,9,18].

4.2. Log-linear functions

Whereas the previous hypotheses were formulated in terms of linear functions of the observed proportions, hypotheses concerned with the analysis of multivariate relationships are sometimes expressed in terms of log-linear models, such as those discussed in Bishop et al. [29]. In particular, hypotheses of "no interaction" in the overall table or in selected marginal distributions discussed in [24,25,30] can all be formulated in terms of functions of the type

$$F(p) = A_2 * log(A_1 * p). \tag{4.3}$$

For further details, see Grizzle and Williams [4]. Other applications of functions in the class of (4.3) include the survival curve analysis of life table data as discussed in Koch, Johnson and Tolley [6] and the analysis of data from paired choice experiments as discussed in Imrey, Koch, and Johnson [12].

4.3. Compounded functions: fixed pattern of transformations

As discussed in Forthofer and Koch [7,22] there are a number of situations in which functions more complex than either (4.2) or (4.3) are required. Among these are the analysis of rank correlation coefficients as discussed in Goodman and Kruskal [31–33] or Davis and Quade [34], the analysis of "ridits" as discussed by Bross [35] or Williams and Grizzle [36], and the analysis of partial association as discussed by Mantel and Haenszel [37] and Mantel [38]. For this purpose, Forthofer and Koch [7] indicated that two general classes of compounded functions could be used to formulate the relevant estimators. They are given as

$$F(p) = A_3 * exp(A_2 * log(A_1 * p)) \tag{4.4}$$

$$F(p) = A_4 * log(A_3 * exp(A_2 * log(A_1 * p))). \tag{4.5}$$

Specific applications of functions of the type in (4.4) and (4.5) are given in [7,22].

4.4. Completely general compounded functions

Although certain patterns of association in square contingency tables can be investigated via functions in (4.5) as discussed in [7,22], there are additional quantities such as complex ratio estimators which cannot be formulated as one of the standard types given in (4.2)–(4.5). For example, the measurement of agreement in multidimensional tables resulting from observer variability studies involves generalized kappa-type statistics of the form

$$F(p) = A_5 * exp(A_4 * log(A_3 * exp(A_2 * log(A_1 * p)))) \tag{4.6}$$

as discussed in Landis [18] and Landis and Koch [19]. Other recent developments utilizing more general forms of compounded functions include the following:

(i) the analysis of life table data using Weibull models as discussed in Freeman, Freeman and Koch [8] using functions of the form

$$F(p) = log(A_2 * log(A_1 * p)); \tag{4.7}$$

(ii) the use of maximum likelihood estimates in fitting hierarchical and non-hierarchical log-linear models by weighted least squares as discussed in Koch, Freeman, Imrey and Tolley [17] using functions of the form

$$F(p) = A_4 * exp(A_3 * log(A_2 * exp(A_1 * p))). \tag{4.8}$$

5. Use of GENCAT

The following sets of cards are used to enter the data and the parameters which determine the type of analysis to be performed:

(0) JOB CONTROL CARDS
(1) BASIC PARAMETER CARD
(2) DATA INPUT CARDS
(3) FUNCTION FORMULATION CARDS

(4) DESIGN MATRIX CARDS
(5) CONTRAST MATRIX CARDS.

Several analyses of a given data set can be performed in the same computer run by fitting more than one design matrix to a particular set of functions and by testing more than one contrast matrix for a particular design. This can be accomplished by simply repeating as many sets of cards (4)–(5) as desired. In addition, the user may specify that a vector and its estimated covariance matrix be saved for reanalysis at the next step. At the reanalysis stage, the saved data are treated as direct input, so that new functions may be defined and analyzed. Within a given sequence of cards (1)–(5), one (and only one) of the following forms of data may be saved for re-analysis – the vector and its estimated covariance matrix corresponding to:

(a) the original data;
(b) the functions resulting from a particular transformation;
(c) the estimated parameters obtained from a given design matrix;
(d) the estimated parameters obtained from a given contrast matrix.

Furthermore, within the same sequence of cards (1)–(5), one (and only one) set of data from (a)–(d) can be punched onto cards (or written to a disk or tape file) to be used in a subsequent computer run. Finally, multiple sets of data can be processed in the same run by simply repeating the appropriate sequence of cards (1)–(5) for each data set.

5.1. Description of the cards

Since the program is written in FORTRAN, all integer-valued parameters must be *right-justified* in their fields on the input cards. All FORMAT statements must be enclosed in parentheses and should be *left-justified* in their fields. Moreover, the input data and matrices must be read according to floating-point specifications involving *F* or *E*, (e.g., 8F10.0, 6E13.5). Fixed-point specifications involving *I* are not permissible.

(0) JOB CONTROL CARDS

These cards are necessary to access and to execute the load module of GENCAT. Because they will vary from one computer system to another, the user will need to determine the specific commands which are required at his/her computer installation. For example, under MTS at the University of Michigan, the required card is as follows: $RUN SGCD:GENCAT 1 = *SOURCE* 3 = *SINK* 8 = –TEMP

Otherwise, at the Trinagle Universities Computation Center, Research Triangle Park, N.C., the required cards for executing the program from a load module stored in UNC.B.F2336.LANDIS.MODKAT are as follows:

```
//GENCAT  JOB UNC.B.XXXXX, USER, T = (,29),
  M = 1, REGION = 200K
//JOBLIB DD DSN = UNC.B.F2336.LANDIS.MODKAT
//  EXEC PGM = GENCAT
//FT08F001 DD DSN =&&TEMP2, DISP = NEW,
//              UNIT = SYSDA, SPACE = (TRK,(1,2)),
//              DCB = (RECFM = VBS, BLKSIZE = 3000)
//FT03F001 DD SYSOUT = A
//FT02F001 DD SYSOUT = B
//FT01F001 DD *
```

(1) BASIC PARAMETER CARD

Columns	*Information contained*
5	Status of data set: 5 = new data; 6 = reanalysis of data saved from previous step.
10	Type of input data (skip if column 5 ≠ 5): 1 = Frequencies from a contingency table: CASE 1 (See Section 3.1); 2 = Frequencies from a contingency table: CASE 2 (See Section 3.2); 3 = Direct input of a function vector and its covariance matrix (See Section 3.3); 4 = Raw data associated with each subject (See Section 3.4).
14–15	Device number from which input data are to be read (skip if column 5 ≠ 5). [e.g., Ø1 = card reader; do not use Ø2, Ø3, or Ø8].

Columns	*Information contained*
25 (optional)	Print options: Ø (or blank) = Print resulting covariance matrix; 1 = Suppress printing of resulting covariance matrix.
30 (optional)	Save options: Ø (or blank) = Do not save initial vector and its covariance matrix for reanalysis; 1 = Save initial vector and its covariance matrix for subsequent analysis in the same run; 2 = Write initial vector and its covariance matrix to unit 2 (typically punched cards).
33–80 (optional)	Title to be printed on first page of analysis.

(2) DATA INPUT CARDS (Skip if column 5 ≠ 5 on (1) BASIC PARAMETER CARD)

The next set of parameter cards is chosen from either (2a), (2b), or (2c) depending on the type of input data indicated in column 10 of the (1) BASIC PARAMETER CARD.

(2a) CONTINGENCY TABLE INPUT CARDS

The cards in this section pertain to input data from contingency tables. Thus, if column 10 ≠ 1 or 2 on the (1) BASIC PARAMETER CARD, skip this section. In CASE 2 the basic block of the diagonal linear operator matrix is read prior to the frequency data. This permits a reduced number of linear functions to be used as the initial vector of proportions. In CASE 1, skip the cards associated with (ii) and (iii).

(i) PARAMETER CARD FOR FREQUENCY DATA (CASE 1 OR 2)

Columns	*Information contained*
1–5	Number of sub-populations (s).
6–10	Number of response profiles (r).
33–80 (optional)	Format by which each row of the contingency table will be read [Default = (8F10.0)].

(ii) PARAMETER CARD FOR BLOCK MATRIX (CASE 2 ONLY)

Colums	*Information contained*
1–5	Number of rows (r') in the basic block of the block diagonal matrix [Be sure $r' * s \leqslant 80$].
33–80 (optional)	Format by which each row of the basic block matrix will be read [Default = (16F5.1)].

(iii) BASIC BLOCK OPERATOR MATRIX (CASE 2 ONLY)

The matrix is entered with each row beginning on a new card according to the format specified either by default or in columns 33–80 of the preceding card (ii).

(iv) FREQUENCY DATA (CASE 1 OR 2)

Regardless of the input device, the contingency table frequencies are entered with each sub-population beginning on a new record according to the format specified either by the default or in columns 33–80 of the (i) PARAMETER CARD FOR FREQUENCY DATA.

(2b) DIRECT INPUT CARDS

The cards in this section pertain to input data in the form of a vector of functions F and its estimated covariance matrix V_F. Thus, if column 10 ≠ 3 on the (1) BASIC PARAMETER CARD, skip this section.

(i) PARAMETER CARD FOR FUNCTIONS (F)

Columns	*Information contained*
1–5	Dimensions of function vector
33–80 (optional)	Format by which the function vector will be read [Default = (5E15.5)].

(ii) VECTOR OF FUNCTIONS (*F*)

Regardless of the input device, the functions are entered as one compound vector (which can usually be partitioned into *s* sub-vectors of dimension *r*) according to the format specified either by the default or in columns 33–80 of the (i) PARAMETER CARD FOR FUNCTIONS (*F*).

(iii) PARAMETER CARD FOR COVARIANCE MATRIX (V_F)

Columns	*Information contained*
1–5	Input mode for covariance matrix 1 = Entire matrix V_F will be read in by rows (each row beginning on a new record); 2 = Upper triangle of V_F will be read in row order as a single vector (each row beginning with the diagonal element of that row); 3 = Diagonal matrix V_F will be read as vector of the diagonal elements.
33–80 (optional)	Format by which V_F will be read [Default = (5E15.5)].

(iv) COVARIANCE MATRIX (V_F)

Regardless of the input device, the matrix is entered according to the format specified either by the default or in columns 33–80 of the (iii) PARAMETER CARD FOR COVARIANCE MATRIX (V_F).

(2c) RAW DATA INPUT CARDS

The cards in this section pertain to input in the form of categorical data variables associated with each subject (See Section 3.4). Thus, if column 10 ≠ 4 on the (1) BASIC PARAMETER CARD, skip this section. (See additional details in Section 5.2).

(i) PARAMETER CARD FOR RAW DATA

Columns	*Information contained*
1–5	Number of independent variables (*q*).
6–10	Number of sub-populations (*s*) to be created on the basis of the (*q*) independent variables.
11–15	Number of dependent variables (*d*).
16–20	Number of functions (*r*) to be created from the (*d*) dependent variables within each sub-population.
33–80 (optional)	Format by which the raw data for each subject will be read [Default = 16F5.0)].

(ii) SUB-POPULATION CARDS

Columns	*Information contained*
1–6	*S*(*JJ*) =, where *JJ* is the sub-population number (If *JJ* is only a single digit, this field ends in column 5 and the next one begins in column 6).
7–80	Statement indicating how the *JJ*-th sub-population is to be formed from the *q* independent variables expressed in terms of operations involving $G(g_1, ..., g_q)$. (See Section 5.2).

(iii) INDICATOR FUNCTION CARDS

Columns	*Information contained*
1–6	*F*(*JJ*) =, where *JJ* is the function number (If *JJ* in only a single digit, this field ends in column 5 and the next one begins in column 6).
7–80	Statement indicating how the *JJ*-th indicator function is to be formed from the *d* dependent variables expressed in terms of operations involving the $G(g_1, ..., g_d)$ and the $W(w_{JJ})$. (See Section 5.2).

(iv) VARIABLE ORDER CARD

This card indicates the order in which the *q* independent (*I*) and *d* dependent (*D*) variables are arranged on the raw data input records. For example, if $q = 2$ and $d = 3$ and the independent variables are in the first and third data fields, this alignment can be denoted by

ORDER = (I,D,I,D,D)

(NOTE: BLANKS ARE NOT PERMITTED WITHIN THE STATEMENT).

Columns	*Information contained*
1–6	ORDER =
7–80	Statement of the form $(I,D,D,I,...,D)$ which indicates the order of the $q + d$ variables on the input records.

(v) DATA CARDS

Regardless of the input device, the data associated with each subject are entered on a new record according to the format specified either by the default or in columns 33–80 of the (i) PARAMETER CARD FOR RAW DATA.

(vi) END OF DATA CARD

The end of the raw input data is indicated by an additional data record which contains a negative integer (e.g., –1) in the first data field.

(3) FUNCTION FORMULATION CARDS

The cards in this section pertain to the formulation of functions from the vector of proportions which was either generated from a contingency table, entered as direct input, calculated from raw data, or saved from a previous step in this same run. These functions are obtained by repeated application of:

(a) linear transformations
(b) logarithmic transformations
(c) exponential transformations
(d) adding a vector of constants.

These transforms can be applied in any order to form the desired compounded functions of the proportions. This is indicated by a series of transformation cards which are ordered according to the particular sequence of application.

(i) TRANSFORMATION CARD

Columns	*Information contained*

5	Type of transformation:
	1 = Linear;
	2 = Logarithmic;
	3 = Exponential;
	4 = Addition of a vector of constants.

10	Input mode for linear operator matrix. (Skip if column 5 ≠ 1): 1 = Entire matrix will be read in by rows; 2 = Basic block of a block diagonal matrix (with identical blocks) will be read in by rows; 3 = Main diagonal of a diagonal matrix will be read in as a vector.
11–15	Number of rows of the linear operator matrix (including all blocks if column 10 = 2). (Skip if column 5 ≠ 1).
16–20	Number of rows in the basic block of the block diagonal matrix. (Skip if column 10 ≠ 2).
25 (optional)	Print options: Ø (or blank) = Print resulting covariance matrix; 1 = Suppress printing of resulting covariance matrix.
30 (optional)	Save options: Ø (or blank) = Do not save resulting vector and its covariance matrix; 1 = Save resulting vector and its covariance matrix for subsequent analysis in the same run; 2 = Write resulting vector and its covariance matrix to unit 2 (typically punched cards).
33–80 (optional)	Format by which each row of the operator matrix (or the corresponding vector if column 5 = 4 or column 10 = 3) will be read [Default = (16F5.1)].

(ii) OPERATOR MATRIX

The cards in this section pertain to the input of a linear operator matrix. Thus, if column 5 ≠ 1 on the (i) TRANSFORMATION CARD, skip this section. The matrix is entered according to the input mode specified in column 10, with each row beginning on a new card according to the format specified either by default or in columns 33–80.

(iii) VECTOR OF CONSTANTS

The cards in this section pertain to the input of a vector of constants. Thus, if column 5 ≠ 4 on the (i) TRANSFORMATION CARD, skip this section. The vector of constants is entered according to the format specified either by default or in columns 33–80.

(4) DESIGN MATRIX CARDS

The cards in this section pertain to a design (or independent variable) matrix X used to investigate the variation among the elements of the function vector F by fitting linear regression models via weighted least squares. This routine will function properly only when the specified set of functions has a non-singular covariance matrix. In addition, the design matrix X must be of full column rank. If a linear model analysis is not desired, skip both sections (4) and (5). Otherwise, the following cards indicate a specific design to be fit to the functions. Any number of design matrices can be fit to a given set of functions (without using the SAVE option) by successively repeating control cards from Sections (4) and (5).

(i) PARAMETER CARD FOR DESIGN MATRIX

Columns	*Information contained*
5	7
10	Input mode for design matrix: 1 = Entire matrix will be read in by columns; 2 = Basic block of a block diagonal matrix (with identical blocks) will be read in by columns; 3 = Main diagonal of a diagonal matrix will be read in as a vector; 4 = Identity matrix.
11–15	Rank (number of columns) of the design matrix.
16–20	Number of columns in the basic block of the block diagonal matrix. (Skip if column 10 ≠ 2.)
25 (optional)	Print options: Ø (or blank) = Print resulting covariance matrix; 1 = Suppress printing of resulting covariance matrix.
30 (optional)	Save options: Ø (or blank) = Do not save resulting vector and its covariance matrix for reanalysis; 1 = Save resulting parameter vector and its covariance matrix for subsequent analysis in the same run; 2 = Write resulting parameter vector and its covariance matrix to unit 2 (typically punched cards).
33–48 (optional)	Format by which each column of the design matrix (or the corresponding vector, if column 10 = 3) will be read. [Default = (16F5.1)].
49–80 (optional)	Title for design matrix.

(ii) DESIGN MATRIX (Skip if column 10 = 4)

The design matrix X is entered according to the input mode specified in column 10 of the preceding (i) PARAMETER CARD with each column beginning on a new card according to the format specified either by default or in columns 33–48.

(5) CONTRAST MATRIX CARDS

The cards in this section pertain to contrast matrices associated with the preceding design matrix specified in Section (4). If no hypotheses involving the parameters in the model are to be tested, skip this section. Otherwise, the following cards are used to test hypotheses of the form $\boldsymbol{C\beta} = \boldsymbol{O}$ for each contrast matrix $\boldsymbol{C}$, where is the vector of model parameters. Any number of hypotheses associated with a particular model can be tested (without using the SAVE option) by successively repeating control cards from this section before fitting another design matrix to the functions.

(i) PARAMETER CARD FOR CONTRAST MATRIX

Columns	*Information contained*
5	8
10	Input mode for contrast matrix: 1 = Entire matrix will be read in by rows; 2 = Basic block of a block diagonal matrix (with identical blocks) will be read in by rows; 3 = Main diagonal of a diagonal matrix will be read in as a vector; 4 = Identity matrix.
11–15	Number of rows of the contrast matrix.
16–20	Number of rows in the basic block of the block diagonal matrix. (Skip if column 10 $\neq$ 2.)
30 (optional)	Save options: ∅ (or blank) = Do not save resulting vector and its covariance matrix; 1 = Save resulting parameter vector and its covariance matrix for subsequent analysis in the same run; 2 = Write resulting parameter vector and its covariance matrix to unit 2 (typically punched cards).
33–48 (optional)	Format by which each row of the contrast matrix (or the corresponding vector, if column 10 = 3) will be read. [Default = (16F5.1)].
49–80 (optional)	Title for contrast matrix.

(ii) CONTRAST MATRIX (Skip if column 10 = 4)

The contrast matrix C is entered according to the input mode specified in column 10 of the preceding (i) PARAMETER CARD with each row beginning on a new card according to the format specified either by default or in columns 33–48.

5.2. Detailed description of raw data control cards

Because the input mode involving raw data is more complex than the other modes discussed in Section 3, this section contains details for the formation of the sub-populations and indicator functions from raw data. The control card instructions presented here are used for (2c) in Section 5.1, only when a particular analysis utilizes raw data. Thus, this section can be bypassed for other modes of data input.

(i) *Specification of sub-populations*

The cards in this section specify how the q independent variables $S_1, S_2, ..., S_q$ are used to form the s sub-populations. For this purpose, the function $G(g_1, g_2, ..., g_q)$ will be used to define the group consisting of those subjects for which $S_i = g_i$ for $i = 1,2,...,q$ (S_1 is the first independent variable on the raw data input cards; S_2 is the second; etc.). These groups must be formed in such a way that each subject fits into exactly one of the sub-population profiles. For example, if $q = 2$ and $S_i = 1,2$, then the sub-populations determined by the four independent variable combinations can be denoted by

$$S(1) = G(1,1)$$
$$S(2) = G(1,2)$$
$$S(3) = G(2,1)$$
$$S(4) = G(2,2).$$

Moreover, subgroups can be formed on the basis of fewer than the q variables by placing a "." in the positions to be ignored. Thus, S_1 can be used to form two sub-populations by setting

$$S(1) = G(1,.)$$
$$S(2) = G(2,.).$$

If all the subjects are to be considered as a random sample from the same sub-population, the grouping array G can be written as

$$S(1) = G(.,.).$$

Sub-populations can also be formed by using "+" to combine various subgroups. For example, if sub-population 1 consists of subjects for which either ($S_1 = 1$ and $S_2 = 2$) or ($S_1 = 2$ and $S_2 = 1$), this can be denoted by

$S(1) = G(1,2) + G(2,1)$.

Each of the s sub-population definitions must begin on a new card. However, the statement can extend to more than one card by simply continuing onto column 1 of the next card. (NOTE: BLANKS ARE NOT PERMITTED WITHIN A STATEMENT).

(ii) *Specification of indicator functions*

The cards in this section specify how the d dependent variables $Y_1, Y_2, ..., Y_d$ will be used to form the u indicator variables. For this purpose, the function $G(g_1, g_2, ..., g_d)$ will be used to denote the response profile in which $Y_i = g_i$ for $i = 1, 2, ..., d$ (Y_1 is the first dependent variable on the raw data input cards; Y_2 is the second; etc.). Furthermore, the weight function $W(w_k)$ will be used to assign the weight w_k to the k-th indicator variable. The default weight function is $w_k = 1.0$, if none is specified.

For example, if $d = 2$ and $Y_i = 1, 2, 3$, then the underlying two-way contingency table proportions can be generated from the following indicator variables (assuming that $w_k = 1.0$ for $k = 2, ..., g$):

$F(1) = G(1,1)$
$F(2) = G(1,2)$
$F(3) = G(1,3)$
$F(4) = G(2,1)$
$F(5) = G(2,2)$
$F(6) = G(2,3)$
$F(7) = G(3,1)$
$F(8) = G(3,2)$
$F(9) = G(3,3)$

If the first order margins are to be estimated, they can be formed by setting:

$F(1) = G(1,.)$
$F(2) = G(2,.)$
$F(3) = G(3,.)$
$F(4) = G(.,1)$
$F(5) = G(.,2)$
$F(6) = G(.,3)$

If these categories are ordinally scaled, an example of an alternative set of indicator functions which can then be used to generate mean scores is given by:

$F(1) = G(1,.) = W(1.0)$
$F(2) = G(2,.) = W(2.0)$
$F(3) = G(3,.) = W(3.0)$
$F(4) = G(.,1) = W(1.0)$
$F(5) = G(.,2) = W(2.0)$
$F(6) = G(.,3) = W(3.0)$

In addition, indicator functions corresponding to sums of the underlying proportions can be specified by using "+", as illustrated by the following main diagonal and first off-diagonal sums which are assigned different weights:

$F(1) = G(1,1) + G(2,2) + G(3,3) = W(1.0)$
$F(2) = G(1,2) + G(2,3) = W(0.5)$
$F(3) = G(2,1) + G(3,2) = W(0.5)$

Each of the u function definitions must begin on a new card. However, the statement can extend to more than one card by simply continuing onto column 1 of the next card. (NOTE: BLANKS ARE NOT PERMITTED WITHIN A STATEMENT.)

6. Examples and sample input cards

In this section we will present several examples of the use of GENCAT with primary attention directed at the preparation of control cards discussed in Section 5. Further details concerning the choice of the appropriate functions and the relevant hypotheses to be tested can be found in the papers cited in the corresponding sections.

6.1. A log-linear model example

This example is based on a research project undertaken at the University of North Carolina Highway Safety Research Center by Stewart [39] for the purpose of studying the relationship between the severity of driver injury in automobile accidents and selected variables characterizing the accident environment with respect to crash configurations, location, time, and weather conditions, automobile type, and driver demographic status. In this regard, the data in table 2 are from a specific, isolated modular component of that investigation which involved the accident sub-population with

Crash configuration	= Single vehicle, medium speed
Location	= Open country
Driver demographic status	= Non-drinking (When accident occurred), male
Calendar year of occurrence	= 1966 or 1968–1972

(6.1)

and its further partition into more refined sub-populations corresponding to the cross-classification of Weather (Good vs. Bad), Time (Day vs. Night), and Model Year (Before 1966 vs. 1967–1969, vs. 1970–1973). The attribute under study is whether or not the driver experienced "severe" injury where "severe" means either an "*A*"-injury (serious visible injury – a bleeding wound, distorted member, or any injury that requires the victim to be carried from the scene) or a "Fatal"-injury (an injury that results in death within 12 months of the accident). Given this framework, the questions of primary statistical interest pertain to the relationship between the conditional probability of "severe" injury and the "Weather," "Time," and "Model Year" characteristics of the accident. For this purpose, one approach of interest is to fit multiple population logistic models to the functions

$$F = F(p) = A_2 * log(A_1 * p), \tag{6.2}$$

where

$$A_1 = I_{24};$$
$$A_2 = [1 -1] \otimes I_{12}, \tag{6.3}$$

and where ⊗ denotes Kronecker product of matrices and I_m is the $m \times m$ identity matrix.

Since preliminary analysis of the data in Table 2 suggested that the second and third order interactions for "Weather", "Time", and "Model Year" were unimportant, the GSK procedure is used to fit the "main effect" model:

$$X = \begin{bmatrix} 1 & 1 & 1 & 1 & 0 \\ 1 & 1 & 1 & 0 & 1 \\ 1 & 1 & 1 & -1 & -1 \\ 1 & 1 & -1 & 1 & 0 \\ 1 & 1 & -1 & 0 & 1 \\ 1 & 1 & -1 & -1 & -1 \\ 1 & -1 & 1 & 1 & 0 \\ 1 & -1 & 1 & 0 & 1 \\ 1 & -1 & 1 & -1 & -1 \\ 1 & -1 & -1 & 1 & 0 \\ 1 & -1 & -1 & 0 & 1 \\ 1 & -1 & -1 & -1 & -1 \end{bmatrix} \tag{6.4}$$

to the logit functions *F* by weighted least squares. Since the goodness of fit statistic $Q = 1.98$ with *D.F.* = 7 is non-significant ($\alpha = 0.25$), the model *X* provides a suitable characterization for these data. In this regard, the estimated parameter vector ***b*** and its estimated covariance matrix V_b are:

Table 2
Tabulation of driver injury by weather, time of day, and model year for 1966, 1968–1972 North Carolina, single vehicle accidents involving non-drinking males and occurring at medium speed in an open country location

Sub-population			Observed frequencies for driver injury	
Weather	Time	Model Year	Not severe	Severe
Good	Day	–1966	5633	898
Good	Day	1967–1969	2371	259
Good	Day	1970–1973	1022	100
Good	Night	–1966	7583	1526
Good	Night	1967–1969	3314	451
Good	Night	1970–1973	1308	168
Bad	Day	–1966	3915	428
Bad	Day	1967–1969	2006	149
Bad	Day	1970–1973	700	43
Bad	Night	–1966	3793	504
Bad	Night	1967–1969	1924	166
Bad	Night	1970–1973	718	51

$$b = \begin{bmatrix} 2.2190 \\ -0.2075 \\ 0.1086 \\ -0.2983 \\ 0.0949 \end{bmatrix};$$

$$V_b = \begin{bmatrix} 5.5506 & & & & \\ -1.1858 & 2.9461 & & \text{(Symmetric)} & \\ 0.3979 & 0.2279 & 2.5355 & & \\ -3.8139 & -0.0141 & -0.0049 & 6.1373 & \\ -1.4226 & 0.1466 & 0.0067 & 0.2028 & 8.6008 \end{bmatrix} \times 10^{-4}. \qquad (6.5)$$

Corresponding Q_C-statistics for testing hypotheses pertaining to b are obtained by using the following C matrices:

$$C_1 = [0 \;\; 1 \;\; 0 \;\; 0 \;\; 0]; \qquad (6.6)$$

$$C_2 = [0 \;\; 0 \;\; 1 \;\; 0 \;\; 0]; \qquad (6.7)$$

$$C_3 = \begin{bmatrix} 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 1 \end{bmatrix}; \qquad (6.8)$$

$$C_4 = [0 \;\; 0 \;\; 0 \;\; 1 \;\; -1]; \qquad (6.9)$$

$$C_5 = [0 \;\; 0 \;\; 0 \;\; 2 \;\; 1]; \qquad (6.10)$$

$$C_6 = [0 \;\; 0 \;\; 0 \;\; 1 \;\; 2]. \qquad (6.11)$$

The sources of variation which correspond to these C matrices and the resulting test statistics are given in table 3. Finally, predicted values $\hat{\pi}_S$ for the conditional probabilities of severe injury based on b can be formulated in terms of compounded functions as

$$\hat{\pi}_S = \pi_S(b) = A_4 * exp(A_3 * log(A_2 * exp(A_1 * b))), \qquad (6.12)$$

where

$$A_1 = \begin{bmatrix} 0.5 \\ -0.5 \end{bmatrix} \otimes X; \quad A_2 = \begin{bmatrix} 0 & 1 \\ 1 & 1 \end{bmatrix} \otimes I_{12};$$

$$A_3 = [1 \;\; -1] \otimes I_{12}; \; A_4 = I_{12}, \qquad (6.13)$$

with the corresponding estimated covariance matrix being determined via (2.5) in conjunction with (2.18)–(2.20). The estimators $\hat{\pi}_S$ based on this approach and their corresponding estimated standard errors are given in the last two columns of table 4. Thus it can be noted that the predicted proportions $\hat{\pi}_S$ are very similar to the original observed proportions (as would be anticipated in view of the acceptable goodness of fit statistic Q) but have substantially smaller estimated standard errors. This gain in statistical efficiency is one of the major advantages of the modeling process.

Table 3
Test statistics for log-linear model effects

Source of variation	D.F.	GSK test statistic
C_1: Weather	1	146.10**
C_2: Time	1	46.52**
C_3: Model year	2	157.79**
C_4: Model year: –1966 vs. 1967–1969	1	107.91**
C_5: Model year: –1966 vs. 1970–1973	1	74.12**
C_6: Model year: 1967–1969 vs. 1970–1973	1	2.84
Residual lack of fit	7	1.98

** means significant at $\alpha = 0.01$

Otherwise, a more complete discussion of the application of weighted least squares methods for fitting logistic and other types of log-linear models is given in Grizzle et al. [1], Grizzle and Williams [4] and Koch et al. [17].

The card preparation necessary for applying GENCAT to example 6.1 is described in the following paragraphs. The input data are frequencies from a contingency table.

(1) BASIC PARAMETER CARD

There are $r = 2$ response profiles within each of $s = 12$ sub-populations, so that $r * s = 24 \leqslant 80$, which allows the data to be entered according to input mode 1. The required parameters and the given run title are shown in fig. 1.

(2) CONTINGENCY TABLE INPUT CARDS

The $r = 2$ frequencies for each sub-population are entered on separate cards according to the default format (8F10.0). Therefore, only the number of sub-populations and the number of response profiles are entered on the parameter card for frequency data which is shown in fig. 1. The twelve data cards containing the observed frequencies from table 2 are entered immediately after the parameter cards shown in fig. 1.

Table 4
Observed and log-linear model predicted proportions of drivers with severe injury for North Carolina data and corresponding standard errors

Weather	Time	Model year	Observed proportion severe injury	Estimated S.E.	GSK log-linear predicted proportion severe injury	Estimated S.E.
Good	Day	–1966	0.1375	0.0043	0.1392	0.0035
Good	Day	1967–1969	0.0985	0.0058	0.0984	0.0035
Good	Day	1970–1973	0.0891	0.0085	0.0892	0.0048
Good	Night	–1966	0.1675	0.0039	0.1673	0.0034
Good	Night	1967–1969	0.1198	0.0053	0.1194	0.0039
Good	Night	1970–1973	0.1138	0.0083	0.1085	0.0055
Bad	Day	–1966	0.0985	0.0045	0.0965	0.0031
Bad	Day	1967–1969	0.0691	0.0055	0.0672	0.0027
Bad	Day	1970–1973	0.0579	0.0086	0.0607	0.0036
Bad	Night	–1966	0.1173	0.0049	0.1172	0.0035
Bad	Night	1967–1969	0.0794	0.0059	0.0822	0.0032
Bad	Night	1970–1973	0.0663	0.0090	0.0744	0.0043

(3) FUNCTION FORMULATION CARDS

The sequence of transformations needed to generate the log-linear function statistics in (6.2) are shown in fig. 2. The logarithmic function is the first to be applied to the proportion vector since the A_1 matrix (6.3) is the identity matrix and therefore has no effect on the functions. The A_2 matrix is a block diagonal matrix and the input mode is specified by a “2” in column 10 of the linear transformation card. The basic block of the A_2 matrix is entered following the corresponding transformation card according to the alternate format of (2F2.0).

(4) DESIGN MATRIX CARDS

The five columns of the design matrix are entered by column as specified by the design matrix parameter card shown in fig. 3. In order that the resulting parameter vector and its covariance matrix be saved for further analysis in the same run, a “1” is placed in column 30 of the parameter card. Each column of the design matrix is entered on a separate card according to the default format (16F5.0). The design matrix follows the parameter card in fig. 3.

(5) CONTRAST MATRIX CARDS

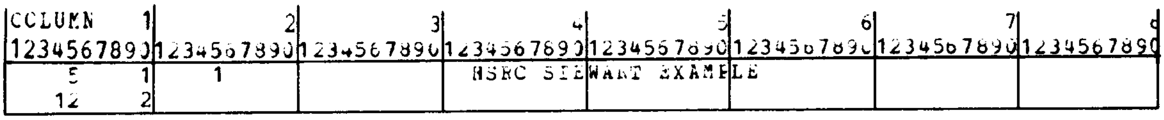
The six contrast matrices given in (6.6)–(6.11) are shown in fig. 4. In each case the matrix is read in by rows according to a format other than the default format. The format specifications begin in column 33 of each contrast matrix parameter card.

(6) REANALYSIS

To obtain predicted values $\hat{\pi}_S$ for the conditional probabilities based on $\boldsymbol{b}$, the parameter vector saved from step (4) is accessed for reanalysis by a second basic parameter card. This card is shown in fig. 5.

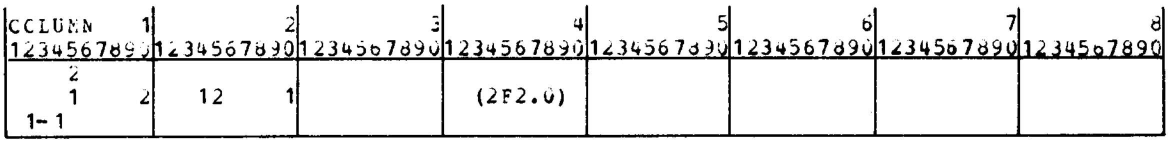
(7) FUNCTION FORMULATION CARDS FOR REANALYSIS

The sequence of transformations required to generate the predicted values $\hat{\pi}_S$ in (6.12) are shown in figs. 5–7. The linear transformation card and the A_1 matrix follow the basic parameter card in fig. 5; the exponential transformation card followed by the linear transformation card and the A_2 matrix are shown in fig. 6; the logarithmic transformation card, the linear transformation card, and the A_3 matrix followed by the final exponential transformation card appear in fig. 7. Since the A_4 matrix is an identity matrix, it



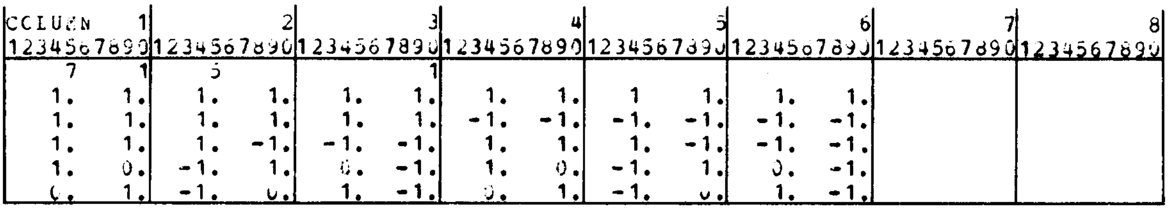
```
CCLUMN   1         2         3         4         5         6         7         8
12345678901234567890123456789012345678901234567890123456789012345678901234567890
    5    1    1                 HSRC STEWART EXAMPLE
   12    2
```

Fig. 1.



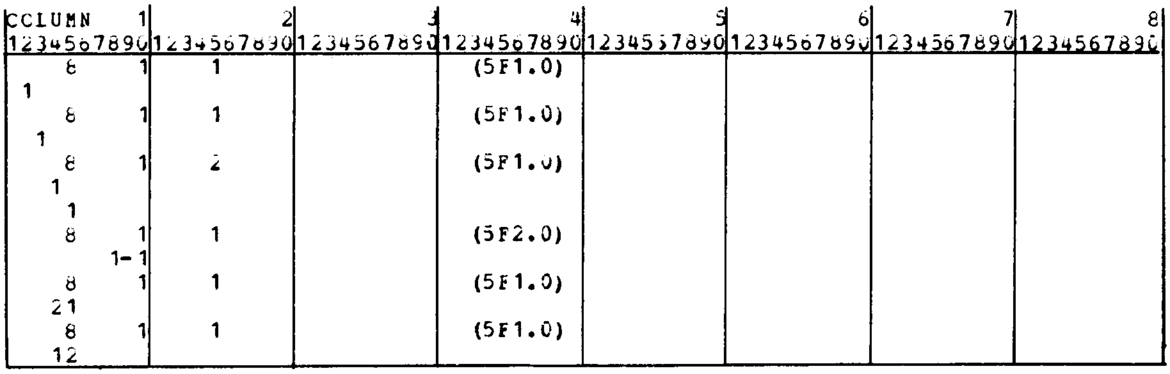
```
CCLUMN   1         2         3         4         5         6         7         8
12345678901234567890123456789012345678901234567890123456789012345678901234567890
    2
    1    2   12    1            (2F2.0)
 1-1
```

Fig. 2.



```
CCLUMN   1         2         3         4         5         6         7         8
12345678901234567890123456789012345678901234567890123456789012345678901234567890
    7    1    5              1
   1.   1.   1.   1.   1.   1.   1.   1.    1   1.   1.   1.
   1.   1.   1.   1.   1.   1.  -1.  -1.  -1.  -1.  -1.  -1.
   1.   1.   1.  -1.  -1.  -1.   1.   1.   1.  -1.  -1.  -1.
   1.   0.  -1.   1.   0.  -1.   1.   0.  -1.   1.   0.  -1.
   0.   1.  -1.   0.   1.  -1.   0.   1.  -1.   0.   1.  -1.
```

Fig. 3.



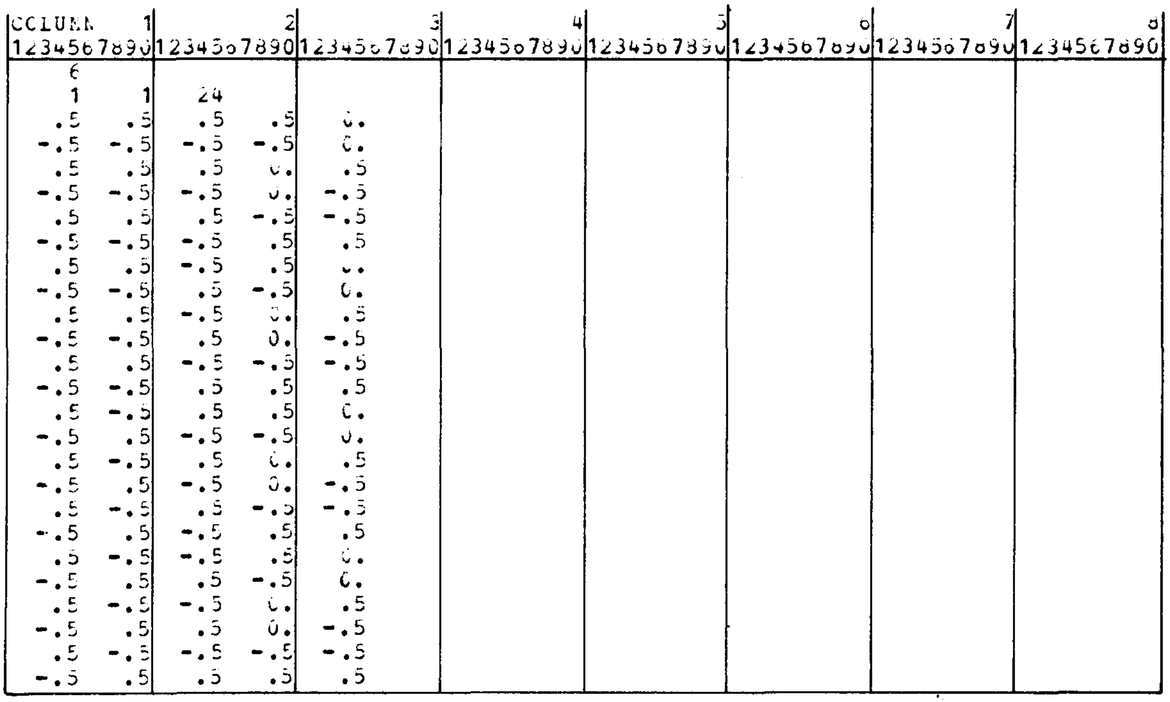
```
CCLUMN   1         2         3         4         5         6         7         8
12345678901234567890123456789012345678901234567890123456789012345678901234567890
    8    1    1                 (5F1.0)
 1
    8    1    1                 (5F1.0)
  1
    8    1    2                 (5F1.0)
   1
    1
    8    1    1                 (5F2.0)
       1-1
    8    1    1                 (5F1.0)
   21
    8    1    1                 (5F1.0)
   12
```

Fig. 4.

need not be entered. Note that the A_2 and A_3 matrices are entered by means of a basic block of a block diagonal matrix, which is indicated by a "2" in column 10 of the corresponding linear transformation parameter cards. The format according to which each linear operator matrix is entered is shown on the respective linear transformation cards (beginning in column 33).

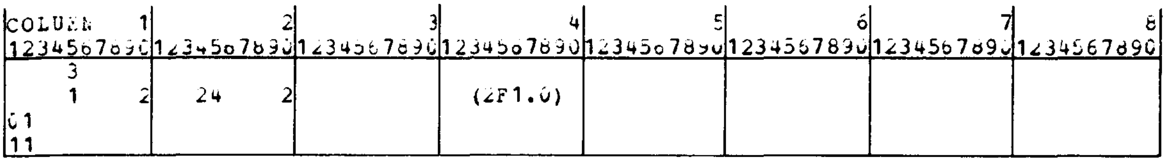
6.2. An observer agreement example

Let us consider the data arising from the diagnosis of multiple sclerosis reported in Westlund and Kurland [40]. Among other things, the investigators were interested in comparing patient groups to study possible differences in the geographical distributions of the dis-



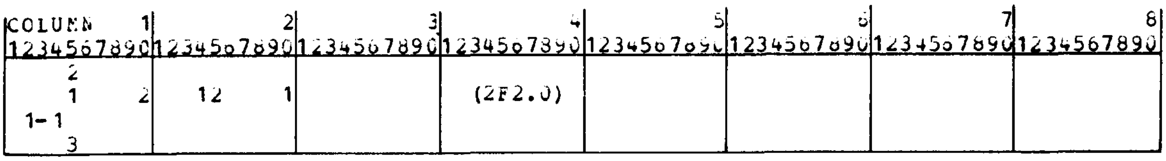
```
CCLUMN   1         2         3         4         5         6         7         8
12345678901234567890123456789012345678901234567890123456789012345678901234567890
    6
    1    1   24
   .5   .5   .5   .5   0.
  -.5  -.5  -.5  -.5   0.
   .5   .5   .5   0.   .5
  -.5  -.5  -.5   0.  -.5
   .5   .5   .5  -.5  -.5
  -.5  -.5  -.5   .5   .5
   .5   .5  -.5   .5   0.
  -.5  -.5   .5  -.5   0.
   .5   .5  -.5   0.   .5
  -.5  -.5   .5   0.  -.5
   .5   .5  -.5  -.5  -.5
  -.5  -.5   .5   .5   .5
   .5  -.5   .5   .5   0.
  -.5   .5  -.5  -.5   0.
   .5  -.5   .5   0.   .5
  -.5   .5  -.5   0.  -.5
   .5  -.5   .5  -.5  -.5
  -.5   .5  -.5   .5   .5
   .5  -.5  -.5   .5   0.
  -.5   .5   .5  -.5   0.
   .5  -.5  -.5   0.   .5
  -.5   .5   .5   0.  -.5
   .5  -.5  -.5  -.5  -.5
  -.5   .5   .5   .5   .5
```

Fig. 5.



```
COLUMN   1         2         3         4         5         6         7         8
12345678901234567890123456789012345678901234567890123456789012345678901234567890
    3
    1    2   24    2            (2F1.0)
01
11
```

Fig. 6.



```
COLUMN   1         2         3         4         5         6         7         8
12345678901234567890123456789012345678901234567890123456789012345678901234567890
    2
    1    2   12    1            (2F2.0)
 1-1
    3
```

Fig. 7.

ease. For this purpose, a series of patients in Winnipeg, Manitoba and a separate series of patients in New Orleans, Louisiana were selected and were examined by a neurologist in their respective locations. After the completion of all the examinations each neurologist was requested to review all the records without seeing his earlier summary and diagnosis, and to classify them into one of the following diagnostic classes:

Table 5
Diagnostic classification regarding multiple sclerosis

Sub-population		Winnipeg patients (1)					
Observer		Winnipeg neurologist (2)					
	Diagnostic class	1	2	3	4	Total	Proportion
New Orleans Neurologist (1)	1	38	5	0	1	44	0.295
	2	33	11	3	0	47	0.315
	3	10	14	5	6	35	0.235
	4	3	7	3	10	23	0.154
	Total	84	37	11	17	149	
	Proportion	0.564	0.248	0.074	0.114		
Sub-population		New Orleans patients (2)					
Observer		Winnipeg neurologist (2)					
	Diagnostic class	1	2	3	4	Total	Proportion
New Orleans Neurologist (1)	1	5	3	0	0	8	0.116
	2	3	11	4	0	18	0.261
	3	2	13	3	4	22	0.319
	4	1	2	4	14	21	0.304
	Total	11	29	11	18	69	
	Proportion	0.159	0.420	0.159	0.261		

1. Certain multiple sclerosis;
2. Probable multiple sclerosis;
3. Possible multiple sclerosis (odds 50:50);
4. Doubtful, unlikely, or definitely not multiple sclerosis.

In order to evaluate agreement between the diagnosticians, the Winnipeg neurologist then reviewed and classified each of the New Orleans patient records, and vice versa. The data resulting from these review diagnoses are presented in table 5.

Although several extensive analyses of these data are discussed in Landis and Koch [19], we will consider only one representative analysis here. Specifically, in order to illustrate the contingency table input mode and the formulation of functions slightly more complex than those illustrated previously in [21,22] we will investigate selected measures of overall agreement between the neurologists. Several questions of interest involve the extent to which the two neurologists classify individual patients into the same diagnostic category. In particular,

(1) Is there any difference between the two patient populations with respect to the overall agreement of the two neurologists on the specific diagnosis of individual patients?

(2) Is the agreement of the two neurologists on the specific diagnosis of individual patients significantly different from chance agreement based on their overall crude distributions of diagnoses?

As stated in [19], these issues can be investigated via kappa-type statistics of the form

$$\hat{\kappa}_{ik} = \frac{\hat{\lambda}_{ik} - \hat{\gamma}_{ik}}{1 - \hat{\gamma}_{ik}}, \tag{6.14}$$

Table 6
Preliminary weights for overall agreement measures

Weights		$w_{1(j)}$				$w_{2(j)}$			
Observer		2				2			
	Diagnostic class	1	2	3	4	1	.2	3	4
Observer 1	1	1	0	0	0	1	1/2	1/4	0
	2	0	1	0	0	1/2	1	1/2	1/4
	3	0	0	1	0	1/4	1/2	1	1/2
	4	0	0	0	1	0	1/4	1/2	1

where $\hat{\lambda}_{ik}$ is an estimate of the observational probability of agreement associated with the k-th set of weights in the i-th sub-population and $\hat{\gamma}_{ik}$ is the corresponding expected proportion of agreement under the baseline constraints of total independence of observer classifications. For this purpose, the weights in table 6 will be used to create preliminary estimates of agreement between the two neurologists. Here $w_{1(j)}$ represents a set of weights selected to generate a measure of perfect agreement, and $w_{2(j)}$ corresponds to a set of weights which assign varying degrees of partial credit to the off-diagonal cells depending on the extent of the disagreement. The estimates of these agreement measures within each of the two patient populations can be expressed in the formulation of (4.6) by choosing:

$$\underset{20\times 32}{A_1} = \begin{bmatrix} 1&1&1&1&0&0&0&0&0&0&0&0&0&0&0&0\\ 0&0&0&0&1&1&1&1&0&0&0&0&0&0&0&0\\ 0&0&0&0&0&0&0&0&1&1&1&1&0&0&0&0\\ 0&0&0&0&0&0&0&0&0&0&0&0&1&1&1&1\\[1ex] 1&0&0&0&1&0&0&0&1&0&0&0&1&0&0&0\\ 0&1&0&0&0&1&0&0&0&1&0&0&0&1&0&0\\ 0&0&1&0&0&0&1&0&0&0&1&0&0&0&1&0\\ 0&0&0&1&0&0&0&1&0&0&0&1&0&0&0&1\\[1ex] 1&0&0&0&0&1&0&0&0&0&1&0&0&0&0&1\\ 1&1/2&1/4&0&1/2&1&1/2&1/4&1/4&1/2&1&1/2&0&1/4&1/2&1 \end{bmatrix} \otimes I_2; \tag{6.15}$$

$$\underset{36\times 20}{A_2} = \begin{bmatrix} 1&0&0&0&1&0&0&0&0&0\\ 1&0&0&0&0&1&0&0&0&0\\ 1&0&0&0&0&0&1&0&0&0\\ 1&0&0&0&0&0&0&1&0&0\\[1ex] 0&1&0&0&1&0&0&0&0&0\\ 0&1&0&0&0&1&0&0&0&0\\ 0&1&0&0&0&0&1&0&0&0\\ 0&1&0&0&0&0&0&1&0&0\\[1ex] 0&0&1&0&1&0&0&0&0&0\\ 0&0&1&0&0&1&0&0&0&0\\ 0&0&1&0&0&0&1&0&0&0\\ 0&0&1&0&0&0&0&1&0&0\\[1ex] 0&0&0&1&1&0&0&0&0&0\\ 0&0&0&1&0&1&0&0&0&0\\ 0&0&0&1&0&0&1&0&0&0\\ 0&0&0&1&0&0&0&1&0&0\\[1ex] 0&0&0&0&0&0&0&0&1&0\\ 0&0&0&0&0&0&0&0&0&1 \end{bmatrix} \otimes I_2; \tag{6.16}$$

$$\underset{8\times 36}{A_3} = \begin{bmatrix} -1&0&0&0&0&-1&0&0&0&0&-1&0&0&0&0&-1&1&0\\ -1&-1/2&-1/4&0&-1/2&-1&-1/2&-1/4&-1/4&-1/2&-1&-1/2&0&-1/4&-1/2&-1&0&1\\ 0&1&1&1&1&0&1&1&1&1&0&1&1&1&1&0&0&0\\ 0&1/2&3/4&1&1/2&0&1/2&3/4&3/4&1/2&0&1/2&1&3/4&1/2&0&0&0 \end{bmatrix} \otimes I_2; \tag{6.17}$$

$$\underset{4\times 8}{A_4} = \begin{bmatrix} 1&0&-1&0\\ 0&1&0&-1 \end{bmatrix} \otimes I_2; \tag{6.18}$$

and

$$\underset{4\times 4}{A_5} = I_4. \tag{6.19}$$

In this context, the A_1 matrix forms the first-order marginal proportions and the weighted sums of observed agreement; the A_2 matrix forms the expected cell proportions by multiplying the appropriate pairs of the observed margins on the $\log_e$ scale; the A_3 matrix forms the corresponding numerator, $\hat{\lambda}_{ik} - \hat{\gamma}_{ik}$, and the denominator, $1 - \hat{\gamma}_{ik}$, for each of the kappa statistics; the A_4 matrix forms the division on the $\log_e$ scale and A_5 selects the required statistics on the anti-$\log_e$ scale. For the data in table 5, these estimates are given by:

$$F = \begin{bmatrix} \hat{\kappa}_{11} \\ \hat{\kappa}_{12} \\ \hat{\kappa}_{21} \\ \hat{\kappa}_{22} \end{bmatrix} = \begin{bmatrix} 0.208 \\ 0.315 \\ 0.297 \\ 0.407 \end{bmatrix}, \tag{6.20}$$

where the $[\hat{\kappa}_{i1}]$ estimate the perfect agreement measure and the $[\hat{\kappa}_{i2}]$ estimate the partial credit weighted agreement measure between the two neurologists in the two patient populations. Using the results in (2.5), the estimated covariance matrix for the estimators in (6.20) is given by

$$V_F = \begin{bmatrix} 0.2546 & 0.2377 & 0.0 & 0.0 \\ 0.2377 & 0.2499 & 0.0 & 0.0 \\ 0.0 & 0.0 & 0.6163 & 0.5623 \\ 0.0 & 0.0 & 0.5623 & 0.5507 \end{bmatrix} \times 10^{-2}. \tag{6.21}$$

The hypotheses associated with questions 1 and 2 can be tested in the linear models phase of the analysis by setting $X = I_4$ and testing each of the following contrast matrices:

$$C_1 = [1 \;\; 0 \;\; 0 \;\; 0]; \tag{6.22}$$

$$C_2 = [0 \;\; 1 \;\; 0 \;\; 0]; \tag{6.23}$$

$$C_3 = [1 \;\; -1 \;\; 0 \;\; 0]; \tag{6.24}$$

$$C_4 = [0 \;\; 0 \;\; 1 \;\; 0]; \tag{6.25}$$

$$C_5 = [0 \;\; 0 \;\; 0 \;\; 1]; \tag{6.26}$$

$$C_6 = [0 \;\; 0 \;\; 1 \;\; -1]; \tag{6.27}$$

$$C_7 = \begin{bmatrix} 1 & 0 & -1 & 0 \\ 0 & 1 & 0 & -1 \end{bmatrix}; \tag{6.28}$$

$$C_8 = [1 \;\; 0 \;\; -1 \;\; 0]; \tag{6.29}$$

$$C_9 = [0 \;\; 1 \;\; 0 \;\; -1]. \tag{6.30}$$

The hypothesis which corresponds to each of these C matrices and the resulting test statistic are given in table 7. These results imply that the perfect agreement measures are not significantly different ($\alpha = 0.25$) and that the weighted agreement measures are not significantly different ($\alpha = 0.25$) in the two groups of patients. Moreover, the various tests of $\kappa_{ik} = 0$ indicate that agreement is significantly ($\alpha = 0.01$) greater than expected under total independence within both groups of patients. However, when compared with each other, the weighted kappa measures are significantly different ($\alpha = 0.01$) from the perfect agreement measures in both sub-populations. This result suggests that the disagreement patterns tended to fall close to the main diagonal (perfect agreement) cells.

Table 7
Statistical tests for agreement statistics using weights from table 6

Hypothesis	*D.F.*	Q_C
Winnipeg patients		
C_1: $\kappa_{11} = 0$	1	16.99**
C_2: $\kappa_{12} = 0$	1	39.70**
C_3: $\kappa_{11} = \kappa_{12}$	1	39.54**
New Orleans patients		
C_4: $\kappa_{21} = 0$	1	14.27**
C_5: $\kappa_{22} = 0$	1	30.07**
C_6: $\kappa_{21} = \kappa_{22}$	1	28.76**
Between sub-populations		
C_7: $\kappa_{11} = \kappa_{21}$; $\kappa_{12} = \kappa_{22}$	2	1.07
C_8: $\kappa_{11} = \kappa_{21}$	1	0.90
C_9: $\kappa_{12} = \kappa_{22}$	1	1.06

**means significant at $\alpha = 0.01$

The following paragraphs contain a detailed description of the card preparation for using GENCAT for this example.

(1) BASIC PARAMETER CARD

Since these data involve $r = 16$ response profiles within each of $s = 2$ sub-populations, $r * s = 32 \leqslant 80$, and thus the data can be entered according to input mode 1. The required parameters and an appropriate run title are shown on the first card in fig. 8.

(2) CONTINGENCY TABLE INPUT CARDS

For this example, the r = 16 frequencies for a given sub-population can be entered on one card by using five-column fields. Because the default format is (8F10.0), the alternative format (16F5.0), must be specified. This parameter card followed by the data cards are also shown in fig. 8.

(3) FUNCTION FORMULATION CARDS

The sequence of transformations required to generate the kappa statistics in (6.20) in the formulation of (4.6) are shown in figs. 9–12. In particular, the linear transformation card and the block matrix for A_1 are shown in fig. 9; the logarithmic transformation card followed by the linear transformation card and the block matrix for A_2 are shown in fig. 10; the exponential transformation card followed by the linear transformation card and the block matrix for A_3 are shown in fig. 11; the logarithmic transformation card, the linear transformation card, the block matrix for A_4 followed by the final exponential transformation card are shown in fig. 12. Note that all the linear transformation cards have a "2" in column 10 to indicate that the matrices are block diagonal.

(4) DESIGN MATRIX CARDS

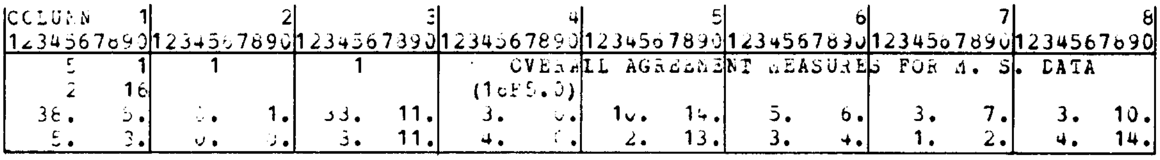
Because the design matrix is set equal to the identity matrix, this can be indicated by one card, which is shown as the first card in fig. 13.

(5) CONTRAST MATRIX CARDS

The contrast matrices given in (6.22)–(6.30) and their corresponding labels are shown in fig. 13.

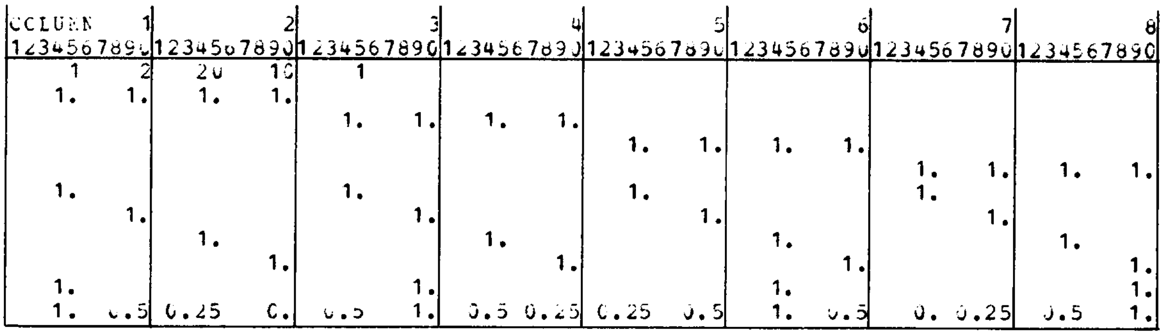
6.3. A raw data example

This example is based on the pathology data reported in [41] which has received extensive analysis recently in Landis and Koch [20]. In order to investigate the variability in the classification of carcinoma in situ of the uterine cervix, seven pathologists were requested to evaluate and to classify 118 slides into one of the following five categories based on the most involved lesion:



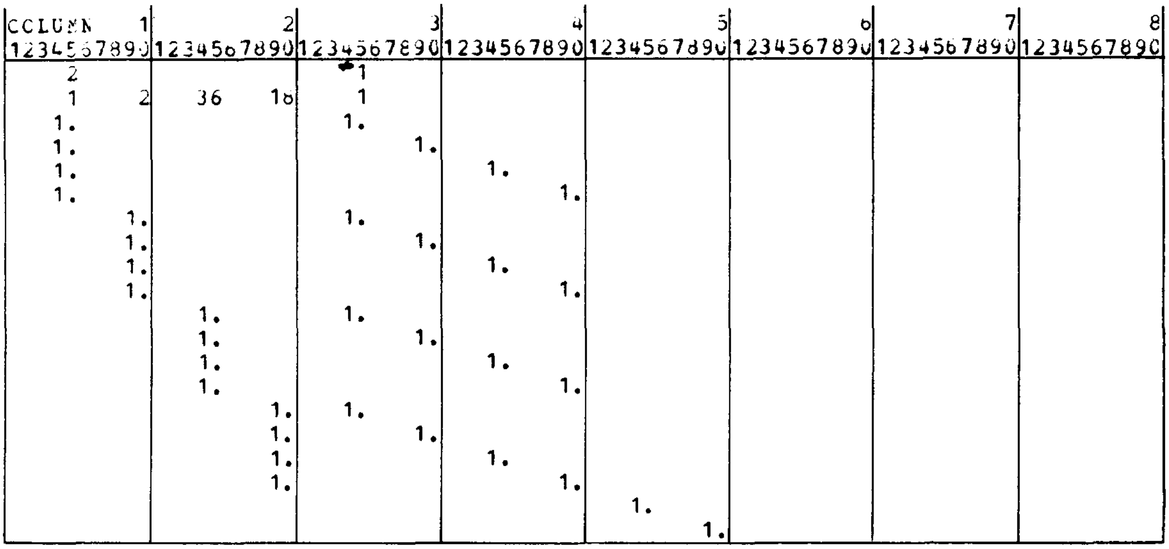
```
COLUMN   1         2         3         4         5         6         7         8
12345678901234567890123456789012345678901234567890123456789012345678901234567890
    5    1    1         1          OVERALL AGREEMENT MEASURES FOR M. S. DATA
    2   16                      (16F5.0)
  38.   5.   0.   1.  33.  11.   3.   0.  10.  14.   5.   6.   3.   7.   3.  10.
   5.   3.   0.   0.   3.  11.   4.   0.   2.  13.   3.   4.   1.   2.   4.  14.
```

Fig. 8.



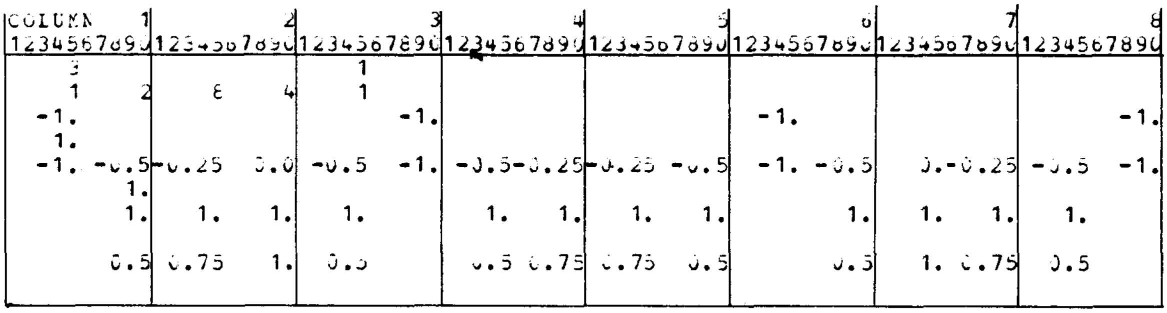
```
COLUMN   1         2         3         4         5         6         7         8
12345678901234567890123456789012345678901234567890123456789012345678901234567890
    1    2   20   10    1
   1.   1.   1.   1.
                       1.   1.   1.   1.
                                           1.   1.   1.   1.
                                                               1.   1.   1.   1.
   1.                  1.                  1.                  1.
        1.                  1.                  1.                  1.
             1.                  1.                  1.                  1.
                  1.                  1.                  1.                  1.
   1.                       1.                       1.                       1.
   1.  0.5 0.25   0.  0.5   1.  0.5 0.25 0.25  0.5   1.  0.5   0. 0.25  0.5   1.
```

Fig. 9.



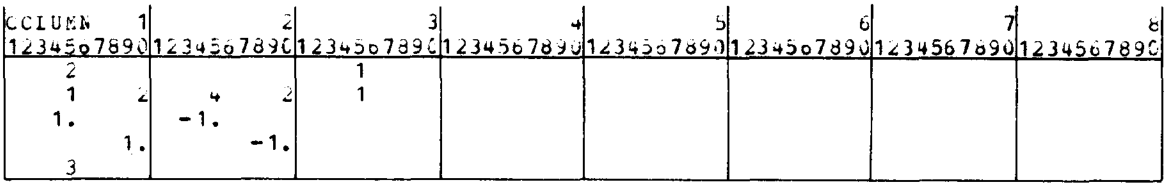
```
COLUMN   1         2         3         4         5         6         7         8
12345678901234567890123456789012345678901234567890123456789012345678901234567890
    2                   1
    1    2   36   16    1
   1.                  1.
   1.                       1.
   1.                            1.
   1.                                 1.
        1.             1.
        1.                  1.
        1.                       1.
        1.                            1.
             1.        1.
             1.             1.
             1.                  1.
             1.                       1.
                  1.   1.
                  1.        1.
                  1.             1.
                  1.                  1.
                                           1.
                                                1.
```

Fig. 10



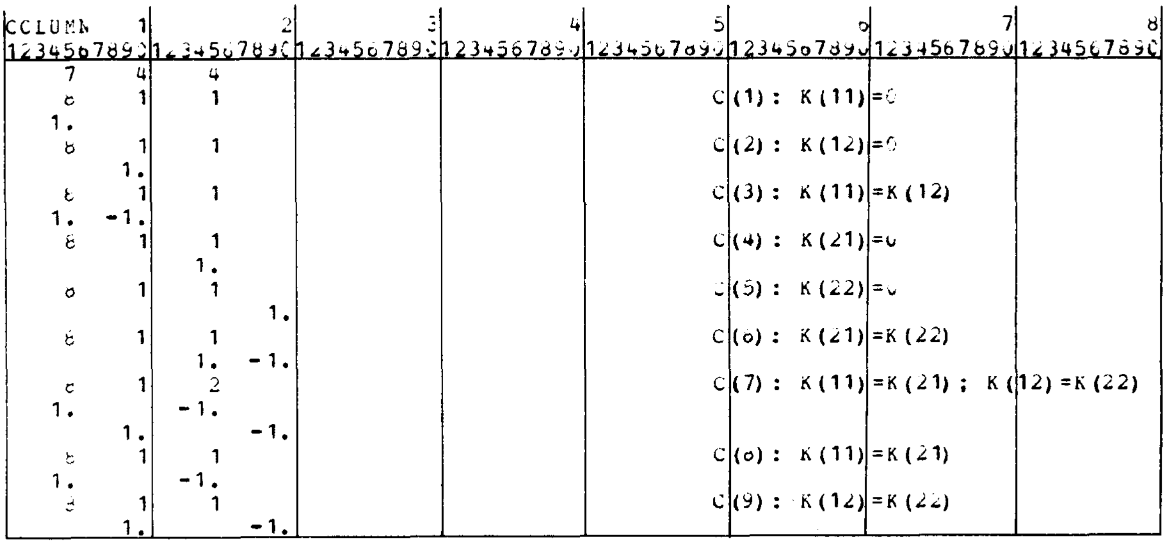
```
COLUMN   1         2         3         4         5         6         7         8
12345678901234567890123456789012345678901234567890123456789012345678901234567890
    3                   1
    1    2    8    4    1
  -1.                      -1.                      -1.                      -1.
   1.
  -1. -0.5-0.25  0.0 -0.5  -1. -0.5-0.25-0.25 -0.5  -1. -0.5   0.-0.25 -0.5  -1.
        1.
        1.   1.   1.   1.        1.   1.   1.   1.        1.   1.   1.   1.

       0.5 0.75   1.  0.5       0.5 0.75 0.75  0.5       0.5   1. 0.75  0.5
```

Fig. 11



```
COLUMN   1         2         3         4         5         6         7         8
12345678901234567890123456789012345678901234567890123456789012345678901234567890
    2                   1
    1    2    4    2    1
   1.       -1.
        1.       -1.
    3
```

Fig. 12



```
COLUMN   1         2         3         4         5         6         7         8
12345678901234567890123456789012345678901234567890123456789012345678901234567890
    7    4    4
    8    1    1                                  C(1) : K(11)=0
   1.
    8    1    1                                  C(2) : K(12)=0
        1.
    8    1    1                                  C(3) : K(11)=K(12)
   1.  -1.
    8    1    1                                  C(4) : K(21)=0
             1.
    8    1    1                                  C(5) : K(22)=0
                  1.
    8    1    1                                  C(6) : K(21)=K(22)
             1.  -1.
    8    1    2                                  C(7) : K(11)=K(21) ; K(12)=K(22)
   1.       -1.
        1.       -1.
    8    1    1                                  C(8) : K(11)=K(21)
   1.       -1.
    8    1    1                                  C(9) : K(12)=K(22)
        1.       -1.
```

Fig. 13

Table 8
Independent classification by seven pathologists of most involved histological lesion

Slide	Pathologist							Slide	Pathologist						
No.	A	B	C	D	E	F	G	No.	A	B	C	D	E	F	G
1	4	3	4	2*	3	3*	3	64	2	3*	2	2	3*	2*	3
2	1	1	1	1	1	1†	1	65	4*	3	3	3	3†	3	3
3	3	3	3	3	3	3	3	66	3	3	3	4*	3†	2	4
4	4*	3*	3	4*	3	3*	3	67	1	1	1	1	1†	1†	1
5	3	3	3	3	3	3*	3	68	2	3*	2	2	3	2	2
6	2*	1	2*	1	1†	1†	1	69	3	3	2	3*	3*	1†	3*
7	1	1	1	1	2*	1†	1	70	1	1	1	1	1†	1†	1
8	3*	3*	2	3*	2*	2	3*	71	4*	3	3	3*	3	3	3
9	2	2*	2	2	3*	1†	2	72	3	3	3	2	3*	1†	3
10	1	1	1	1	2†	1†	1	73	3	3	3	2	3	2	3
11	5	5	5	4	5	5	5	74	4	3	1	3	3	2	3
12	1	1	1	1	2†	1†	1	76	1	2*	1	1	1†	1†	1
13	3	3	3	2	3	3*	3	77	2	2*	1	2	2	1†	2*
15	2	2*	2	1	1†	1†	2	78	2	3	2	1	3	2	2
16	4*	3	3	2	3	2*	3	79	2	1*	1	2	1†	1†	1
17	3	3	2	3	3	3	3	80	4	4*	3	2	4*	1*	3
18	2*	3*	2	2	3	2	3*	81	1	1	1	1	1†	1†	1
19	2	1*	2	1	2†	1†	1	82	4	4*	3	3	4*	3	3
22	2	3*	2	2*	2	1†	3*	83	5	5	1	4	5	5	4*
23	1	1*	2	1	1	1†	1	84	2	3	2	2	2*	1†	2
24	4*	3	3	4*	3	3	3	85	4	4*	4*	2	5	1*	3
25	1	1	2	1	2†	1†	1	86	3	3	2	3	3	3*	3
26	1	1	1	1	1†	1†	1	87	4	3	3	3*	3	3	3
27	2	1	2	2	2*	1†	2	88	4*	2	3	2	3	2*	3
28	4	4*	4*	2*	4	3	3	89	2	3	2	2	4	1†	3
29	3	3	3	2	3	2	3*	90	3	3	3	2	4*	2	3
30	3	3	3	3*	3	2	3	91	3*	3	2	1*	3	2	2
31	1	1	1	1	1†	1†	1	92	4	4*	3	2	4*	1*	3
32	4*	3	3	3*	3*	2*	3†	93	3*	3	2†	2	3	2*	2*
33	3	3	3	3	3	3	3†	94	1	1	2†	1	2*	1†	1
34	1	1	1	1	1†	1†	1	95	3	3	3	2*	4	3	3
35	3	3	3	2	3	1†	3	96	4*	3*	1	1	2*	1†	2
36	2	2*	2	2	3*	1†	2	98	4*	3	3	4	4	3	2
37	3	3	2	2*	3	1†	3	99	1	2*	2	1	2*	1†	2*
38	3*	3	3	3*	4	1†	3	100	3	3	3	2	4*	2	3
39	2	1	1	1	2	1†	1	101	4	4*	3	4	4	3	4*
40	3	3	2	2	3	1†	3†	102	3	3	2	2*	3*	3*	3*
41	3	3	3	3*	3	2	3*	103	1	1	1	1	1†	1†	1
42	5	5	5	5	5	5	5	104	2	3*	2	2	4*	1†	2
43	5	3	3	2*	3	2	3	105	3	3	3	3*	3	2	3
44	3*	2*	2*	2	2*	1†	2	106	2	3	1	1	3*	1†	1
45	1	1	1	1	2	1†	1	107	3	3	2	2*	3	2	3*
46	2	3*	1	2	3*	1†	3*	108	3	3	2	2	3	1†	3*
47	4*	4*	4	3*	3	3	3	110	2	2*	1	1	2	1†	1
48	3	3	3	2	3	2	3	111	1	1†	1	1	2	1†	1
49	3*	2*	2	2	2*	1†	1	112	3	3	2	2	2*	2	3
51	2	3*	2	2†	2	2	2	113	3	3*	2*	2	2	1†	2
52	3	3	3	4*	3	2*	3	114	2	3*	1	1	2*	1†	1
53	4*	3	3	3	3	5*	3	115	3	3*	2	2	3	2	3

Table 8 (continued)

Slide	Pathologist							Slide	Pathologist						
No.	A	B	C	D	E	F	G	No.	A	B	C	D	E	F	G
54	3	3	2	2	4*	2	3*	116	1	1	1	1	2	1†	1
55	3	3	3*	3	3	2*	3	117	3	3	3	2	3	2	3
56	2	2*	2	1	2*	2	2*	118	3*	3*	2*	2	3*	1†	3*
57	2	3	2	2	3	1†	3*	119	1	1	1	1	2	1†	1
58	1	1	1	1	1†	1†	1	120	1	1*	1	1	1	1†	1
59	3*	3	3*	3	3	3	3*	121	2	2*	1	1	2	1†	2
60	1	1	2	1	1†	1†	1	122	5*	3*	4	2	3	4*	3
61	1	3*	2	1	2	1†	1	123	4	3	4	2	4*	1†	3
62	4*	3	3	3*	3	2	3	124	1	1	1	1	2	1†	1
63	1	3	2	2	2	1†	2	126	2	3*	1	1	2*	1†	2

* Indicates doubtful classification.
† Indicates no statement of confidence.

1. Negative;
2. Atypical squamous hyperplasia;
3. Carcinoma in situ;
4. Squamous carcinoma with early stromal invasion;
5. Invasive carcinoma.

This particular design involves $s = 1$ sub-population, $d = 7$ observers, and $L = 5$ response categories which produces $r = L^d = 78,125$ possible response profiles. Obviously, with only $n = 118$ slides most of the corresponding cell frequencies in the underlying multidimensional contingency table are zero. Furthermore, the sizes of the operator matrices associated with the direct GSK approach to the analysis of these data are outside the scope of computational feasibility. However, indicator functions of the raw data in table 8 can be formulated to investigate the relationships of interest as discussed in Section 3.4.

Although several statistical issues in these data are investigated in [20], we will consider only one representative analysis here to illustrate the raw data input mode and the utilization of indicator functions. In particular, the question concerning interobserver bias in the overall usage of the diagnostic scale can be addressed in terms of hypotheses of first-order marginal homogeneity. If these diagnostic categories are indexed by $j_g = 1, 2, ..., 5$ for each of the pathologists indexed by $g = 1, 2, ..., 7$, then the indicator variables which can be used to estimate the corresponding marginal probabilities can be denoted by

$$z_{kgl} = \begin{cases} 1, & \text{if the } k\text{-th pathologist classifies the } l\text{-th slide into the } g\text{-th category} \\ 0, & \text{otherwise.} \end{cases} \qquad (6.31)$$

Thus, these functions can be expressed in the notation of (3.3) by letting

$$\underset{1\times 35}{z_l'} = (z_{11l}, ..., z_{15l}, z_{21l}, ..., z_{25l}, ..., z_{71l}, ..., z_{75l}) \qquad (6.32)$$

As a result, the estimates of the marginal probabilities for the diagnostic classes "1", "2", "3", "4", "5" for each of the seven pathologists can be obtained as across-slide arithmetic means via (3.2). These estimates are displayed in table 9.

Table 9
First-order margins of seven pathologists classifying 118 slides according to most involved histological lesion

Pathologist	Response category				
	1	2	3	4	5
A	0.220	0.220	0.322	0.186	0.051
B	0.229	0.102	0.585	0.059	0.025
C	0.263	0.356	0.314	0.051	0.017
D	0.322	0.407	0.195	0.068	0.008
E	0.136	0.263	0.449	0.119	0.034
F	0.525	0.263	0.169	0.008	0.034
G	0.271	0.169	0.517	0.025	0.017

Table 10
Combined classes for two-point scale

Class	Original classification
C_1	1, 2
C_2	3, 4, 5

Table 11
Statistical tests of marginal homogeneity using two-point scale

Q_C Values for pairwise tests between pathologists (d.f. = 1)

Pathologist	A	B	C	D	E	F	G
A	–	9.62	25.55	47.76	1.49	62.83	0.00
B		–	44.12	78.11	4.76	99.56	12.46
C			–	7.17	33.35	16.25	25.55
D				–	54.19	2.64	47.76
E					–	75.39	2.32
F						–	62.83
G							–

In view of substantial differences reflected in these marginal distributions in table 9 the diagnostic criteria for the five point scale do not appear to be sufficiently precise to ensure a high level of interobserver agreement. Consequently, the effects of a reduction in the scale shown in table 10 will be investigated. This dichotomous classification is of considerable clinical importance, since different types of follow-up may be prescribed for patients diagnosed as C_1 than for those diagnosed as C_2. In this situation, the marginal proportions corresponding to C_1 for each pathologist could be generated from the vector of means associated with (6.32) by letting

$$\underset{7\times 35}{A_1} = [1 \quad 1 \quad 0 \quad 0 \quad 0] \otimes I_7. \tag{6.33}$$

Nevertheless, in order to illustrate the derivation of these functions directly from the raw data, let

$$z_{kl} = \begin{cases} 1, & \text{if the } l\text{-th slide is classified as } C_1 \text{ by} \\ & \text{pathologist } k \\ 0, & \text{otherwise.} \end{cases} \tag{6.34}$$

Then the mean vector associated with (6.34) for the data in table 8 is given by

$$F = \bar{z} = \begin{bmatrix} 0.441 \\ 0.331 \\ 0.619 \\ 0.729 \\ 0.398 \\ 0.788 \\ 0.441 \end{bmatrix} \tag{6.35}$$

which contains the estimates of the probability of assignment to C_1 by each pathologist. The pairwise hypotheses of marginal homogeneity can be tested by letting $X = I_7$ and by using C matrices of the form

$$C_{hh'} = [c_1, c_2, c_3, c_4, c_5, c_6, c_7], \tag{6.36}$$

where

$$c_k = \begin{cases} 1, & \text{if } k = h \\ -1, & \text{if } k = h' \\ 0, & \text{otherwise.} \end{cases} \tag{6.37}$$

The resulting test statistics associated with these contrast matrices are given in table 11. Otherwise, a more complete discussion of the analysis of these data is given in [20].

The card preparation necessary for using GENCAT to obtain the analyses of the data in table 8 is given in the following paragraphs.

(1) BASIC PARAMETER CARD

Since there are $d = 7$ dependent variables and $q = 1$ independent variable (here the entire sample is considered as $s = 1$ sub-population) in the form of integer-valued variables, these data are entered according to input mode 4. The required parameters and the given run title are shown on the first card in fig. 14.

(2) RAW DATA INPUT CARDS

(i) PARAMETER CARD FOR RAW DATA

The required parameters to generate the $u = 35$ indicator functions in (6.31) and the appropriate format statement are shown on the second card in fig. 14.

(ii) SUB-POPULATION CARD

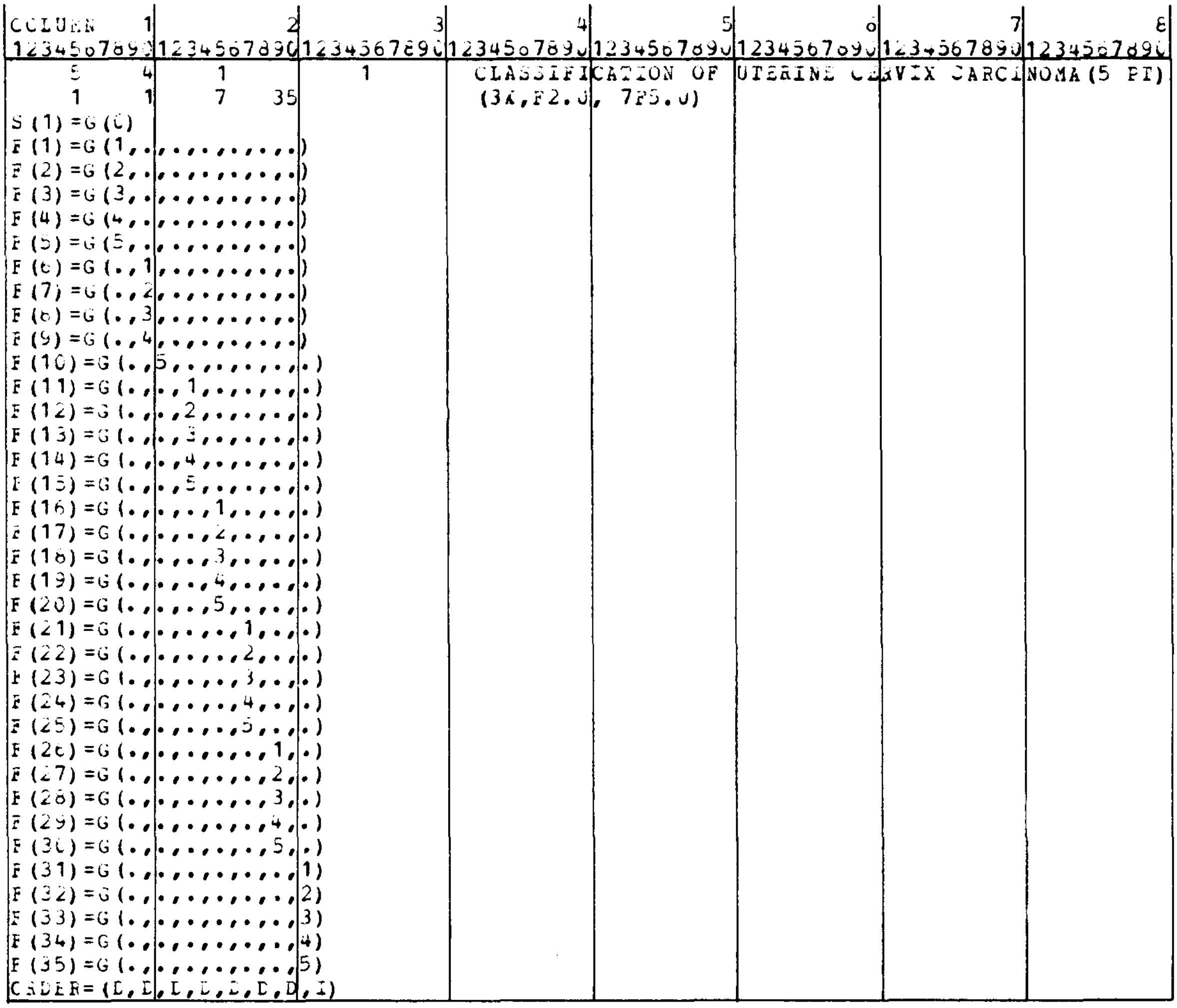
Since there is only one sub-population, i.e., the entire sample is assumed to be from the same population, a "dummy" independent variable is read from a blank field on each card to create $S_1 = 0$ for each slide. Thus, the appropriate statement to indicate that all the slides are from the same sub-population is shown on the third card in fig. 14.

(iii) INDICATOR FUNCTION CARDS

The $u = 35$ indicator variables specified in (6.31) can be generated by corresponding function cards shown in fig. 14. Since these functions involve first-order margins, the sums over the other variables are indicated by a "." in the appropriate positions of the grouping function G as discussed in Section 5.2.

(iv) VARIABLE ORDER CARD

The card indicating the order of the $d = 7$ dependent and $q = 1$ independent variables on the data cards is shown in fig. 14 immediately after the indicator function cards.



```
COLUMN   1         2         3         4         5         6         7         8
12345678901234567890123456789012345678901234567890123456789012345678901234567890
    5    4 1            1       CLASSIFICATION OF UTERINE CERVIX CARCINOMA(5 PT)
    1    1 7      35            (3X,F2.0, 7F5.0)
S(1)=G(0)
F(1)=G(1,.,.,.,.,.,.)
F(2)=G(2,.,.,.,.,.,.)
F(3)=G(3,.,.,.,.,.,.)
F(4)=G(4,.,.,.,.,.,.)
F(5)=G(5,.,.,.,.,.,.)
F(6)=G(.,1,.,.,.,.,.)
F(7)=G(.,2,.,.,.,.,.)
F(8)=G(.,3,.,.,.,.,.)
F(9)=G(.,4,.,.,.,.,.)
F(10)=G(.,5,.,.,.,.,.)
F(11)=G(.,.,1,.,.,.,.)
F(12)=G(.,.,2,.,.,.,.)
F(13)=G(.,.,3,.,.,.,.)
F(14)=G(.,.,4,.,.,.,.)
F(15)=G(.,.,5,.,.,.,.)
F(16)=G(.,.,.,1,.,.,.)
F(17)=G(.,.,.,2,.,.,.)
F(18)=G(.,.,.,3,.,.,.)
F(19)=G(.,.,.,4,.,.,.)
F(20)=G(.,.,.,5,.,.,.)
F(21)=G(.,.,.,.,1,.,.)
F(22)=G(.,.,.,.,2,.,.)
F(23)=G(.,.,.,.,3,.,.)
F(24)=G(.,.,.,.,4,.,.)
F(25)=G(.,.,.,.,5,.,.)
F(26)=G(.,.,.,.,.,1,.)
F(27)=G(.,.,.,.,.,2,.)
F(28)=G(.,.,.,.,.,3,.)
F(29)=G(.,.,.,.,.,4,.)
F(30)=G(.,.,.,.,.,5,.)
F(31)=G(.,.,.,.,.,.,1)
F(32)=G(.,.,.,.,.,.,2)
F(33)=G(.,.,.,.,.,.,3)
F(34)=G(.,.,.,.,.,.,4)
F(35)=G(.,.,.,.,.,.,5)
ORDER=(D,D,D,D,D,D,D,I)
```

Fig. 14

(v) DATA CARDS

The data for each slide in table 8 was punched on a separate card according to the format shown in fig. 15. For illustrative purposes only the first and last five cards of this data file are given here.

(vi) END OF DATA CARD

The appropriate card to indicate the end of the input data file is shown as the last card in fig. 15.

Thus, these cards shown in figs. 14 and 15 are sufficient to generate the marginal distributions shown in table 9. Moreover the estimated covariance matrix for these functions is also computed, so that the further modeling and hypotheses testing discussed in [20] can be performed in the usual GSK framework.

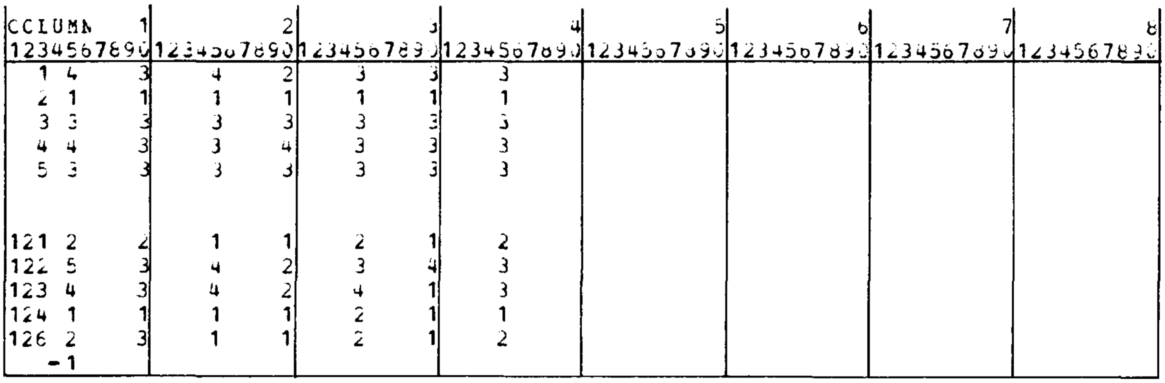
However, attention is now directed at the additional analysis of these data utilizing the two-point scale presented in table 10. Accordingly, the appropriate cards used to generate the $r = 7$ indicator functions in (6.34) in a separate computer run are shown in fig. 16. These cards are then followed by the same input data cards shown in fig. 15. Since the function vector F in (6.35) is to be analyzed directly, no cards from the Section (3) FUNCTION FORMULATION CARDS are required. Thus, these cards in figs. 16 and 15 are followed immediately by the design and contrast matrix cards.

(4) DESIGN MATRIX CARDS

Because the design matrix is set equal to the identity matrix, this can be indicated by one card, which is shown as the first card in fig. 17.

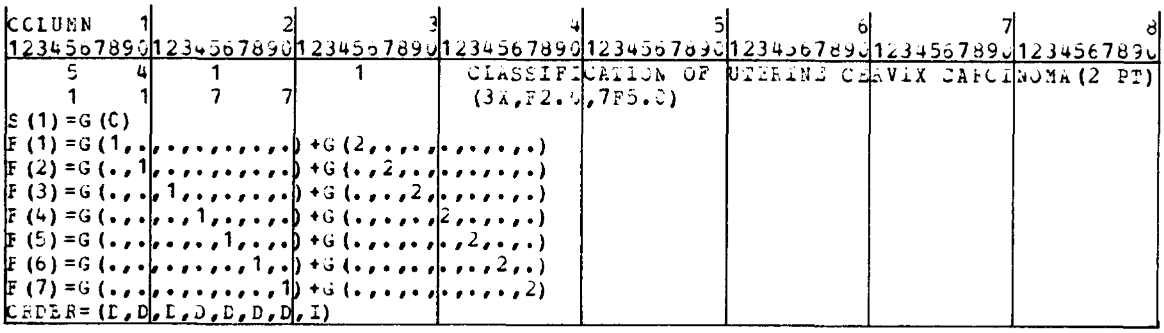
(5) CONTRAST MATRIX CARDS

The contrast matrices specified in (6.36) and (6.37)



```
CCLUMN   1         2         3         4         5         6         7         8
12345678901234567890123456789012345678901234567890123456789012345678901234567890
  1 4    3    4    2    3    3    3
  2 1    1    1    1    1    1    1
  3 3    3    3    3    3    3    3
  4 4    3    3    4    3    3    3
  5 3    3    3    3    3    3    3


121 2    2    1    1    2    1    2
122 5    3    4    2    3    4    3
123 4    3    4    2    4    1    3
124 1    1    1    1    2    1    1
126 2    3    1    1    2    1    2
   -1
```

Fig. 15



```
CCLUMN   1         2         3         4         5         6         7         8
12345678901234567890123456789012345678901234567890123456789012345678901234567890
    5    4    1         1      CLASSIFICATION OF UTERINE CERVIX CARCINOMA(2 PT)
    1    1    7    7           (3X,F2.0,7F5.0)
S(1)=G(C)
F(1)=G(1,.,.,.,.,.,.)+G(2,.,.,.,.,.,.)
F(2)=G(.,1,.,.,.,.,.)+G(.,2,.,.,.,.,.)
F(3)=G(.,.,1,.,.,.,.)+G(.,.,2,.,.,.,.)
F(4)=G(.,.,.,1,.,.,.)+G(.,.,.,2,.,.,.)
F(5)=G(.,.,.,.,1,.,.)+G(.,.,.,.,2,.,.)
F(6)=G(.,.,.,.,.,1,.)+G(.,.,.,.,.,2,.)
F(7)=G(.,.,.,.,.,.,1)+G(.,.,.,.,.,.,2)
CRDER=(D,D,D,D,D,D,D,I)
```

Fig. 16

and their corresponding labels are shown in fig. 17.

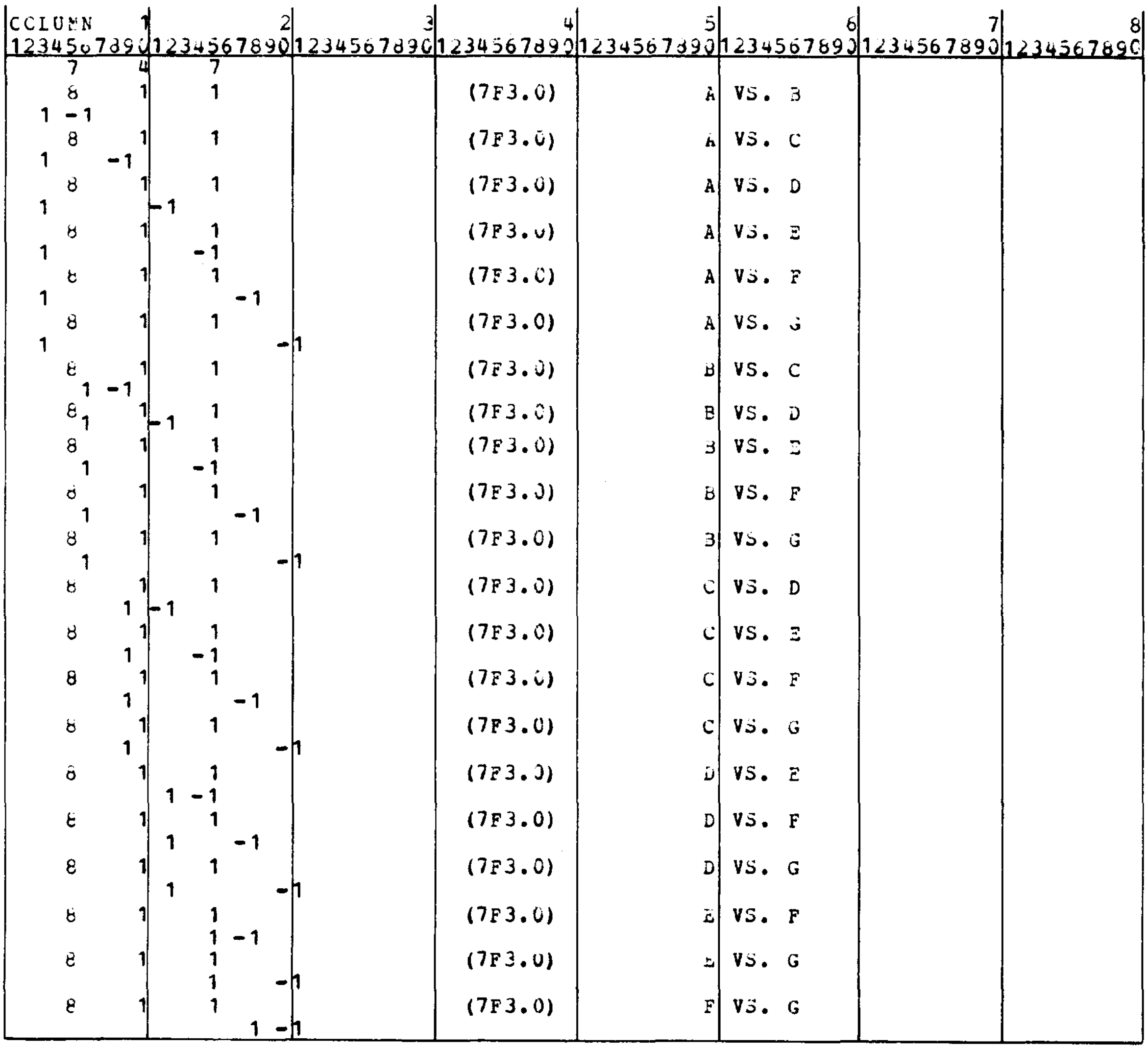
7. Description of error messages

1. "THE NUMBER OF FUNCTIONS EXCEEDS (R – 1) * S = ——, · · · IN THIS CASE INDUCING A SINGULAR COVARIANCE MATRIX. YOU WILL NEED A LINEAR OPERATOR WITH —— ROWS OR LESS. PROGRAM TERMINATING."

This message is printed if the user has input frequency data and has more than $(r-1)*s$ functions when he is done with the transformation stage of the program (note that there is a maximum of $(r-1)*s$ linearly independent functions of the original $r*s$ proportions).

2. "CANNOT REANALYZE DATA · · · NOTHING SAVED. PROGRAM ENDING."

This message is printed if the user has specified reanalysis of the same data but did not save any data with a "save option".



```
CCLUMN   1         2         3         4         5         6         7         8
12345678901234567890123456789012345678901234567890123456789012345678901234567890
    7    4    7
    8    1    1                 (7F3.0)          A VS. B
  1 -1
    8    1    1                 (7F3.0)          A VS. C
  1    -1
    8    1    1                 (7F3.0)          A VS. D
  1       -1
    8    1    1                 (7F3.0)          A VS. E
  1          -1
    8    1    1                 (7F3.0)          A VS. F
  1             -1
    8    1    1                 (7F3.0)          A VS. G
  1                -1
    8    1    1                 (7F3.0)          B VS. C
     1 -1
    8    1    1                 (7F3.0)          B VS. D
     1    -1
    8    1    1                 (7F3.0)          B VS. E
     1       -1
    8    1    1                 (7F3.0)          B VS. F
     1          -1
    8    1    1                 (7F3.0)          B VS. G
     1             -1
    8    1    1                 (7F3.0)          C VS. D
        1 -1
    8    1    1                 (7F3.0)          C VS. E
        1    -1
    8    1    1                 (7F3.0)          C VS. F
        1       -1
    8    1    1                 (7F3.0)          C VS. G
        1          -1
    8    1    1                 (7F3.0)          D VS. E
           1 -1
    8    1    1                 (7F3.0)          D VS. F
           1    -1
    8    1    1                 (7F3.0)          D VS. G
           1       -1
    8    1    1                 (7F3.0)          E VS. F
              1 -1
    8    1    1                 (7F3.0)          E VS. G
              1    -1
    8    1    1                 (7F3.0)          F VS. G
                 1 -1
```

Fig. 17

3. "ERROR IN DATA - - - PROGRAM TERMINATED."

This message is printed if the user specifies an invalid numeric code - - - for example, a "4" on the save option, or an "8" on the suppress-print option.

4. "RANK OF BASIC BLOCK MUST BE POSITIVE - - - PROGRAM ENDING."

This is printed if rank $(A^*) < 1$ and the user has input frequency data, case 2.

5. "THE NUMBER OF ROWS (COLUMNS) OF THE ENTIRE MATRIX MUST BE GREATER THAN THE NUMBER OF ROWS (COLUMNS) OF THE BASIC BLOCK. IF EQUALITY HOLDS, THERE IS ONLY ONE BLOCK IN THE MATRIX, IN WHICH CASE THERE SHOULD BE A 1 IN COLUMN 10 OF THE PARAMETER CARD DESCRIBING THIS MATRIX. PROGRAM TERMINATING."

6. "TOO MANY SUBPOPULATION CARDS - - - PROGRAM TERMINATING."

In the raw data option, the user specifies on the parameter card the number of sub-populations to be formed. This message is printed if the number of sub-population cards exceeds this specification.

7. "NOT ENOUGH SUB-POPULATION CARDS - - - PROGRAM TERMINATING."

This message is printed if the number of sub-population cards is less than the number specified on the parameter card.

8. "ERROR - - - K'TH SUB-POPULATION (OR FUNCTION) CARD IS NOT LABELED AS SUCH. PLEASE RELABEL. PROGRAM TERMINATING. FAULTY CARD:" ——————————."

This message is printed when the user fails to comply with the arbitrary restriction that the K'th sub-population (or function) card begin with "S(K)= " ("F(K)= ").

9. "INVALID INPUT. PROGRAM TERMINATING. FAULTY CARD: " ——————————."

This message is printed (when using the raw data option) if input is invalid - - - either because of what is on the card, or because of the placement of the card within the deck.

10. "INVALID INPUT IN RAW DATA PARAMETER CARD - - - PROGRAM ENDING."

An example of invalid input would be a negative number of sub-populations.

11. "ERROR - - - THE NUMBER OF VALUES INSIDE THE PARENTHESES SHOULD BE ——. THE NUMBER FOUND WAS ——. INVALID INPUT. PROGRAM TERMINATING. FAULTY CARD: ——————————."

For sub-population cards (in the raw data option) the number of values inside the grouping parentheses $[G(1, 2, 1)]$ should be q - - - the number of independent variables. For function cards, the number should be d - - - the number of dependent variables. This message is printed if there is an incorrect number of variables within the grouping parentheses.

12. "ERROR - - - INCORRECT NUMBER OF FUNCTION CARDS - - - THERE SHOULD BE ——. PROGRAM TERMINATING."

This message is printed (when using the raw data option) if the number of function cards does not agree with the number of functions specified on the raw data parameter card.

13. "ERROR - - - SUB-POPULATION PROFILE NOT FOUND FOR THE FOLLOWING DATA VECTOR:" ——————————."

A sub-population profile is the set of q-tuples (corresponding to the values of the q independent variables) such that if a subject has a raw data vector matching one of the q-tuples in this set, then it is placed in that sub-population. (The purpose of each sub-population card is to define its profile.) This error message is printed when a data vector is found for which there is no matching sub-population profile.

14. "WARNING: YOU SPECIFIED THAT THE PRECEDING VECTOR AND ITS COVARIANCE MATRIX BE SAVED. IT IS NOT SAVED BECAUSE THERE IS A PREVIOUS ENTRY IN THE SAVE AREA FOR THIS PASS."

A "pass" is defined to start with a basic parameter card and to end with the next basic parameter card. Within *each pass*, one vector and covariance matrix may be saved internally and one may be punched (or written to an appropriate file). This warning is printed if there is an attempt to save more than one internally or to write more than one to an appropriate file, within one pass.

8. Hardware specifications

This computer program 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 in order to use this program on other machines. However, a revised version of GENCAT is currently being prepared which will permit the program to run on a wide range of machines.

9. Program availability

The source deck for both the batch and time-sharing versions of GENCAT, together with the corresponding listings, may be obtained for a nominal cost from the Department of Biostatistics, School of Public Health, University of Michigan, Ann Arbor, Michigan, 48109, U.S.A. A current version of the documentation and running instructions is included with the initial purchase of the computer program.

Purchasers may choose to keep their names on an active mailing list to receive information concerning updating and further modifications of the program. For instance, work is currently underway to develop an additional option which automatically creates standard design and contrast matrices associated with the usual ANOVA parameterizations of main effects and interactions. Implementing this capability will involve adding one additional subroutine to the existing source code. Persons on the mailing list will be notified when such revisions are available for distribution.

10. Disclaimer

Although GENCAT has received extensive testing, 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 questions or problems do arise, contact the first author at the Department of Biostatistics, School of Public Health, University of Michigan, Ann Arbor, Michigan, 48109, U.S.A.

Acknowledgments

The authors wish to thank numerous graduate students and research assistants in the Departments of Biostatistics at the University of North Carolina and at the University of Michigan who provided many helpful comments during the debugging of various test versions of GENCAT. In addition, Eugene R. Heyman, Joel D. Kaufman, Alcinda L. Lewis, and Karen A. Stanecki were helpful in the card preparation for the examples, Robert J. Anderson provided valuable assistance in constructing figs. 1–17, and David F. Hunsche prepared the final form of the flowcharts shown in the Appendix. Moreover, the authors are grateful to Dr. Dwight B. Brock of the National Center for Health Statistics and Dr. Daniel H. Freeman, Jr. of Yale University for many useful comments in the development of the program documentation. Finally, the authors gratefully acknowledge the untiring efforts of Ms. Linda L. Blakley, Ms. Laura A. Grice and Ms. Cindy M. Coon in typing the original draft of this manuscript, and the assistance of Ms. Therese L. Birney in typing the final revisions.

References

[1] J.E. Grizzle, C.F. Starmer and G.G. Koch, Biometrics 25 (1969) 489–504.

[2] W.D. Johnson and G.G. Koch, Technometrics 13 (1971) 438–447.

[3] G.G. Koch and D.W. Reinfurt, Biometrics 27 (1971) 157–173.

[4] J.E. Grizzle and O.D. Williams, Biometrics 28 (1972) 137–156.

[5] G.G. Koch, P.B. Imrey and D.W. Reinfurt, Biometrics 28 (1972) 663–692.

[6] G.G. Koch, W.D. Johnson and H.D. Tolley, J. Amer. Statist. Assoc. 67 (1972) 783–796.

[7] R.N. Forthofer and G.G. Koch, Biometrics 29 (1973) 143–157.

[8] D.H. Freeman, Jr., J.L. Freeman and G.G. Koch, University of North Carolina Institute of Statistics Mimeo Series No. 958 (1974), J. Amer. Statist. Assoc., submitted for publication.

[9] G.G. Koch, J.L. Freeman, D.H. Freeman, Jr., and R.G. Lehnen, University of North Carolina Institute of Statistics Mimeo Series No. 961 (1974), Biometrics, submitted for publication.

[10] N. El Khorazaty, University of North Carolina Institute of Statistics Mimeo Series No. 1019 (1975), Ph.D. dissertation.

[11] D.H. Freeman, Jr., University of North Carolina Institute of Statistics Mimeo Series No. 1020 (1975), Ph.D. dissertation.

[12] P.B. Imrey, G.G. Koch and W.D. Johnson, University of North Carolina Institute of Statistics Mimeo Series No. 1004 (1975), J. Amer. Statist. Assoc., to appear.

[13] G.G. Koch, D.H. Freeman, Jr. and J.L. Freeman, Internat. Statist. Rev. 43 (1975) 55–74.

[14] G.G. Koch, J.L. Freeman and R.G. Lehnen, Internat. Statist. Rev. 44 (1976) 1–28.

[15] G.G. Koch and H.D. Tolley, Biometrics 31 (1975) 59–92.

[16] G.G. Koch, H.D. Tolley and J.L. Freeman, University of North Carolina Institute of Statistics Mimeo Series No. 864 (1975), Biometrics, to appear.

[17] G.G. Koch, D.H. Freeman, Jr., P.B. Imrey and H.D. Tolley, University of North Carolina Institute of Statistics Mimeo Series No. 1046 (1975), Biometrics, submitted for publication.

[18] J.R. Landis, University of North Carolina Institute of Statistics Mimeo Series No. 1022 (1975), Ph.D. dissertation.

[19] J.R. Landis and G.G. Koch, Biometrics, to appear.

[20] J.R. Landis and G.G. Koch, Biometrics, to appear.

[21] R.N. Forthofer, C.F. Starmer and J.E. Grizzle, J. Biomed. Sys. 2 (1971) 3–48.

[22] R.N. Forthofer and G.G. Koch, Comp. Prog. in Biomed. 3 (1974) 237–248.

[23] A. Wald,Trans. Amer. Math. Soc. 54 (1943) 426–482.

[24] V.P. Bhapkar and G.G. Koch, Technometrics 10 (1968) 107–123.

[25] V.P. Bhapkar and G.G. Koch, Biometrics 24 (1968) 567–594.

[26] J. Neyman, in: Proceedings of the Berkeley Symposium of Mathematical Statistics and Probability (Univ. of California Press, Berkeley and Los Angeles, 1949), pp. 239–273.

[27] V.P. Bhapkar, J. Amer. Statist. Assoc. 61 (1966) 228–235.

[28] N. Nie, C.H. Hull, J.G. Jenkins, K. Steinbrenner, D. Brent, Statistical Package for the Social Sciences (McGraw Hill, 1975).

[29] Y.M.M. Bishop, S.E. Fienberg, P.W. Holland, Discrete Multivariate Analysis (M.I.T. Press, 1975).

[30] S.N. Roy and M.A. Kastenbaum, Ann. Math. Statist. 27 (1956) 749–757.

[31] L.A. Goodman and W.H. Kruskal, J. Amer. Statist. Assoc. 49 (1954) 732–764.

[32] L.A. Goodman and W.H. Kruskal, J. Amer. Statist. Assoc. 54 (1959) 123–163.

[33] L.A. Goodman and W.H. Kruskal, J. Amer. Statist. Assoc. 58 (1963) 310–364.

[34] C.E. Davis and D. Quade, Biometrics 24 (1968) 987–996.

[35] I.D.J. Bross, Biometrics 14 (1958) 18–38.

[36] O.D. Williams and J.E. Grizzle, J. Amer. Statist. Assoc. 67 (1972) 55–63.

[37] N. Mantel and W. Haenszel, J. Nat. Cancer Inst. 22 (1959) 719–748.

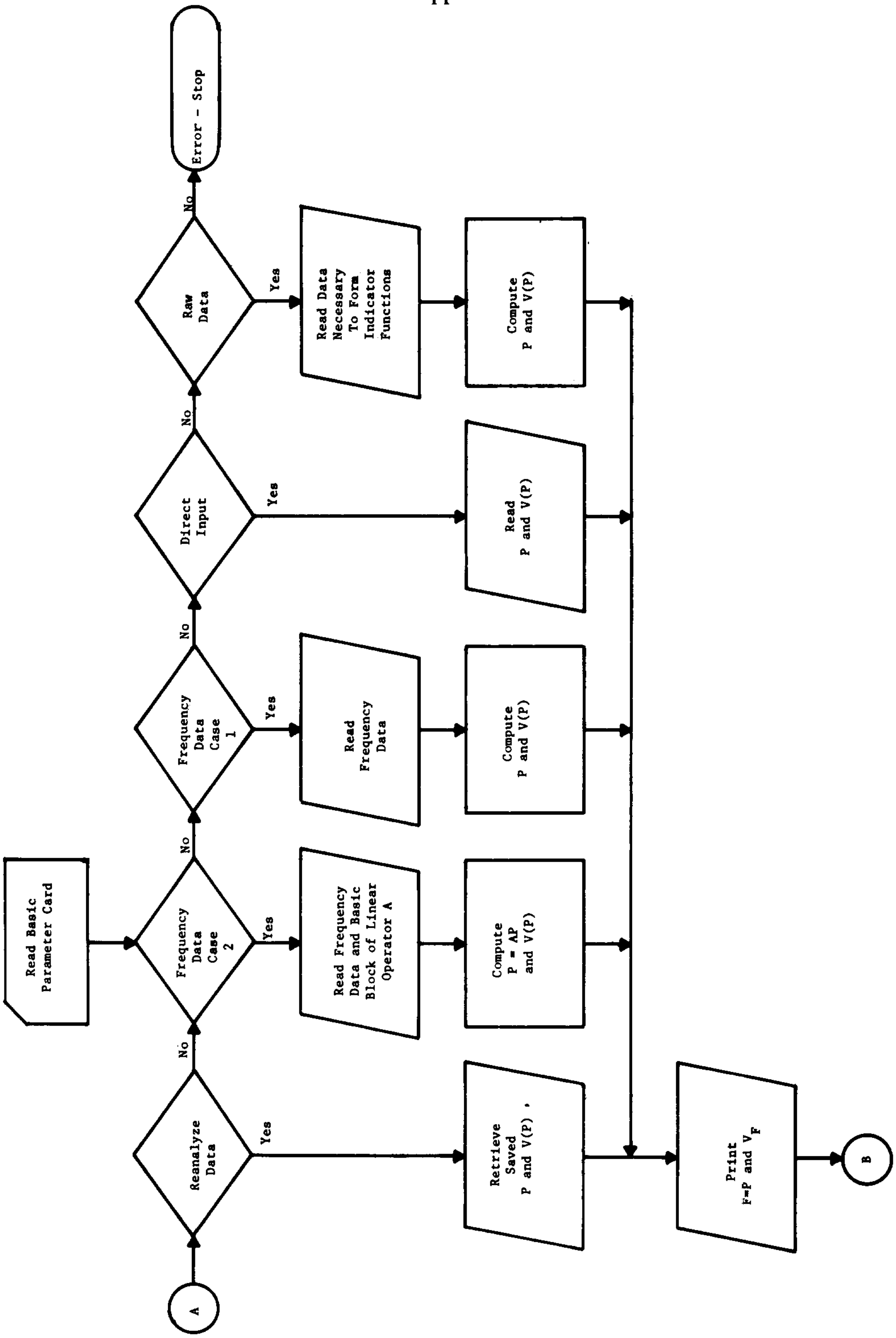
[38] N. Mantel, J. Amer. Statist. Assoc. 58 (1963) 690–700.

[39] J.R. Stewart, University of North Carolina Highway Safety Research Center Technical Report (1975).

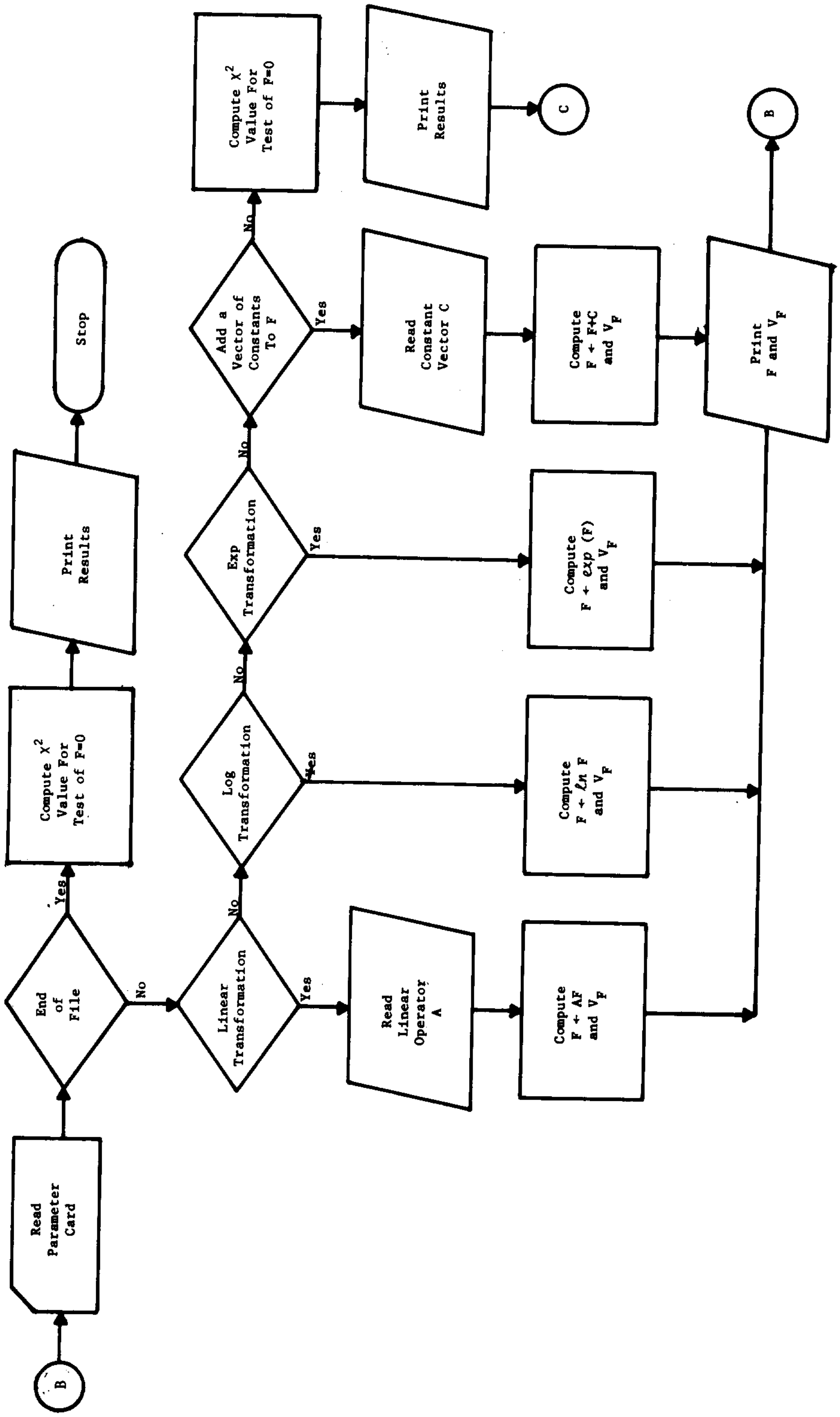
[40] K.B. Westlund and L.T. Kurland, Am. J. Hyg. 57 (1953) 380–396.

[41] N.D. Holmquist, C.A. McMahan and O.D. Williams, Arch. Path. 84 (1967) 334–345.

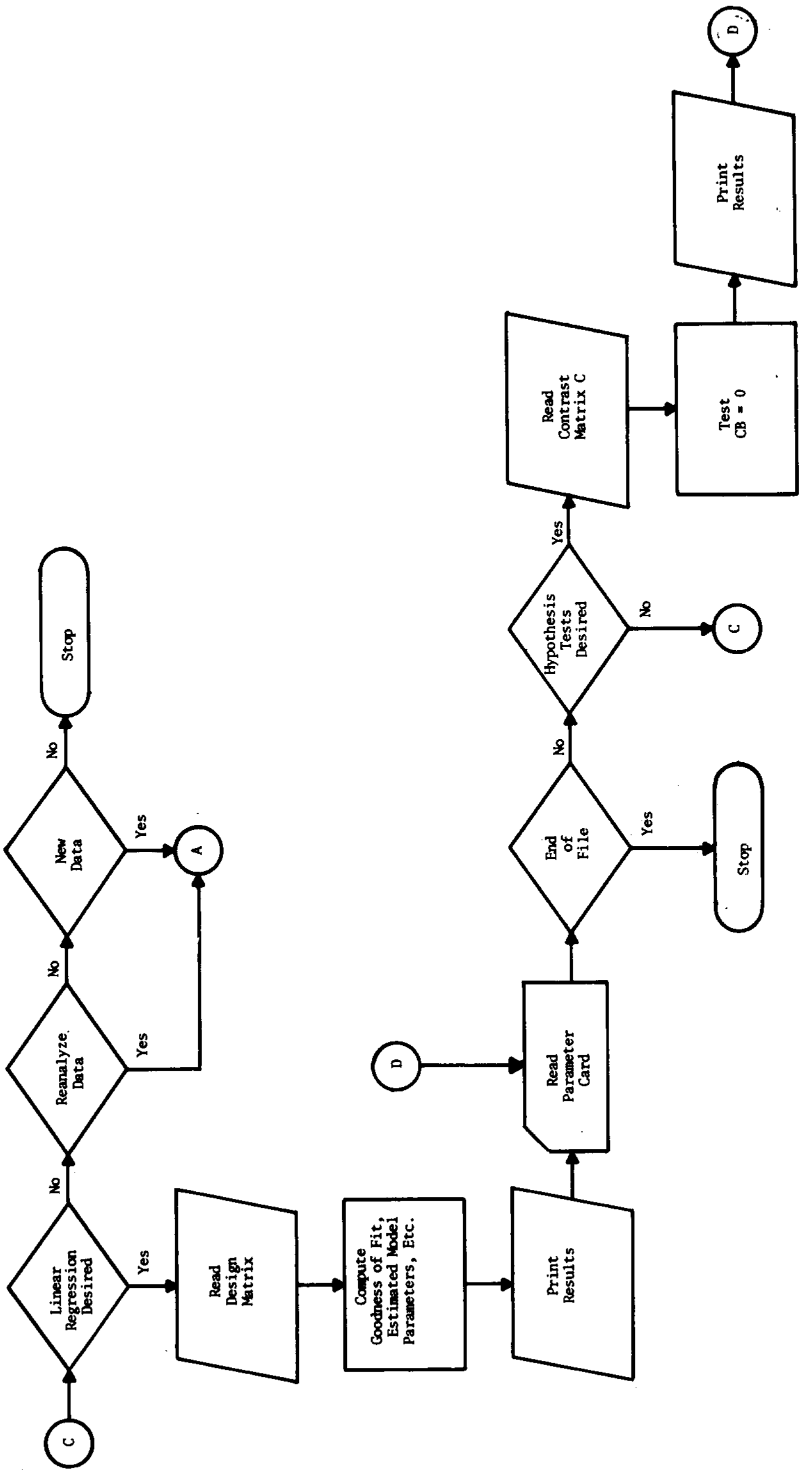
Appendix 1


A
Read Basic Parameter Card
Reanalyze Data
Yes
No
Frequency Data Case 2
Yes
No
Frequency Data Case 1
Yes
No
Direct Input
Yes
No
Raw Data
Yes
No
Error - Stop
Retrieve Saved P and V(P)
Read Frequency Data and Basic Block of Linear Operator A
Compute P = AP and V(P)
Read Frequency Data
Compute P and V(P)
Read P and V(P)
Read Data Necessary To Form Indicator Functions
Compute P and V(P)
Print F=P and V_F
B

Appendix 2


B
Read Parameter Card
End of File
Yes
No
Compute χ^2 Value For Test of F=0
Print Results
Stop
Linear Transformation
Yes
No
Read Linear Operator A
Compute F ← AF and V_F
Log Transformation
Yes
No
Compute F ← ℓn F and V_F
Exp Transformation
Yes
No
Compute F ← exp (F) and V_F
Add a Vector of Constants To F
Yes
No
Read Constant Vector C
Compute F ← F+C and V_F
Print F and V_F
B
Compute χ^2 Value For Test of F=0
Print Results
C

Appendix 3


C
Linear Regression Desired
Yes
No
Read Design Matrix
Compute Goodness of Fit, Estimated Model Parameters, Etc.
Print Results
Read Parameter Card
D
End of File
Yes
Stop
No
Hypothesis Tests Desired
No
C
Yes
Read Contrast Matrix C
Test CB = 0
Print Results
D
Reanalyze Data
Yes
No
New Data
Yes
A
No
Stop