Open access • Journal Article • DOI:10.2307/2528755

Linear model analysis of categorical data with incomplete response vectors.

Gary G. Koch, Peter B. Imrey, Donald W. Reinfurt

Published on: 01 Sep 1972 - Biometrics (Wiley-Blackwell)

Topics: Categorical variable, Missing data, Linear model, General linear model and Contingency table

This research was supported by the National Institutes of Health, Institute of General Medical Sciences Grants GM-70004-01, GM-0038-18, and GM-12868-08.

LINEAR MODEL ANALYSIS OF CATEGORICAL DATA WITH INCOMPLETE RESPONSE VECTORS

by

Gary G. Koch,[1] Peter B. Imrey,[1] and Donald W. Reinfurt[2]

[1]Department of Biostatistics and [2]Highway Safety Research Center
University of North Carolina at Chapel Hill

Institute of Statistics Mimeo Series No. 790

December 1971

SUMMARY

The general linear model approach to the analysis of categorical data described by Grizzle, Starmer and Koch [1969] is extended to situations where: (1) missing data for certain individuals arise at random as a result of non-response or deleted incorrect response; (2) supplemental samples pertaining to various subsets of variables have been obtained due to cost considerations and/or special interest in these variables. The problems discussed are distinct from those involving "incomplete contingency tables" containing a priori empty cells.

The extension is presented through a series of examples which show how the approach can be used to handle a wide variety of non-standard data configurations. Applications to categorical data mixed models and split plot designs are emphasized.

LINEAR MODEL ANALYIS OF CATEGORICAL DATA WITH INCOMPLETE RESPONSE VECTORS

Gary G. Koch[1], Peter B. Imrey[1], and Donald W. Reinfurt[2]
[1]Department of Biostatistics and [2]Highway Safety Research Center
University of North Carolina, Chapel Hill, N.C. 27514, U.S.A.

1. INTRODUCTION

In a recent paper, Grizzle, Starmer, and Koch [1969] (subsequently abbreviated GSK) described how linear regression models and weighted least squares could be used to analyze multivariate categorical data which has been summarized in a multi-dimensional contingency table. This methodology provides the researcher with a powerful and flexible tool which can be used to evaluate models and to test hypotheses for a broad class of experimental situations. For example, various conditions of "no interaction" as discussed by Roy and Kastenbaum [1956] and Bhapkar and Koch [1968a, 1968b] can be investigated either as they apply to cell probabilities or to certain functions thereof (e.g., marginal probabilities, logits, mean scores, etc.). The resulting test statistics belong to the class of minimum modified - χ^2_1, due to Neyman [1949], and are equivalent to certain generalized quadratic form criteria of Wald [1943]. Alternatively, many of these same problems can be approached by using the methods described by Mantel [1966], Lewis [1968], Bishop [1969,1971], Fienberg [1970], and Goodman [1970, 1971a, 1971b] based on maximum likelihood, or those described by Ku, Varner, and Kullback [1968, 1971] based on minimum discrimination information.

All of the previously mentioned papers are primarily concerned with data where there is complete classification of experimental units (or subjects) with respect to all variables of interest. However, in certain types of applications incomplete configurations are encountered either due to chance or design. Examples here include:

1. Cases where missing data arise as a result of non-response or deleted incorrect response, but such missing data can be presumed to occur at random.
2. Cases where supplemental samples pertaining to a subset of variables have been obtained either because of greater interest in those variables or economic cost considerations. These situations are analogous to those which arise in the context of augmented fractional factorial designs as discussed by Box and Wilson [1951], Box [1966], John [1966], Gaylor and Merrill [1968], and sequences of fractional factorials as discussed by Daniel [1962] and Addelman [1969]; this relationship is particularly apparent when one is concerned with a categorical data mixed model or split plot experiment as described by Koch and Reinfurt [1970].
3. Incomplete block or fractional factorial type split plot multivariate designs like those described by Roy, Gnanadesikan, and Srivastava [1971].
4. Growth curve experiments like those described by Potthoff and Roy [1964] and Allen and Grizzle [1969].

The principal objective of the analysis for any of these situations is to obtain valid and precise estimators for all of the relevant parameters in a manner which uses as much of the available information in the data as possible (i.e.,subjects with incomplete responses are not necessarily omitted). With respect to incomplete categorical data, Hocking and Oxspring [1971] have discussed maximum likelihood for the case of sampling from a single multinomial population in terms of an approach similar to that given by Hocking and Smith [1968] for incomplete multivariate normal data. A similar problem has been considered by Blumenthal [1968], also in terms of maximum likelihood. Alternatively, Koch and Reinfurt [1970] have described a two-stage procedure which yields the

minimum-χ^2_1 analysis of incomplete categorical data. The remainder of this paper will be concerned with describing some additional aspects of this extension of the GSK approach and with illustrating its application to several pertinent examples.

Finally, it is important to note that the situations under consideration here are entirely different from those corresponding to "incomplete contingency tables"as discussed by Goodman [1968], Bishop and Fienberg [1969], Williams and Grizzle [1970], and Mantel [1970]. In these cases, each experimental unit is classified according to all of the variables of interest, but certain combinations of such variables have zero probability of occurrence. Thus, the resulting contingency table is incomplete in the sense that certain cells contain zero frequencies regardless of sample size. Hence, the term "incomplete contingency table" is a suitable descriptive title for such situations. However, the question then arises as to what to call the incomplete categorical data situations of interest in the remainder of this paper. For lack of a better term cases like those cited in (1) or (2) will be referred to as "augmented contingency tables" while the ones like those cited in (3) or (4) will be called "complex split plot contingency tables" depending on the nature of the experimental situation. Some of the reasoning behind these descriptive titles will be apparent from the nature of the examples which will be considered.

2. SOME RELEVANT EXAMPLES

2.1 An application with missing data (non-response)

In Table 1, data are given for the unaided distance vision of 8577 women aged 30-39. For 7477 of these women, vision grade was classified for both eyes and the resulting frequencies are identical to the data used as an illustrative example for tests of marginal homogeneity by several authors

TABLE 1

UNAIDED DISTANCE VISION; 8577 WOMEN AGED 30-39

Right Eye	Left eye: Highest Grade (1)	Second Grade (2)	Third Grade (3)	Lowest Grade (4)	Sub-Total	Right Only	Total
Highest grade (1)	1520	266	124	66	1976	140	2116
Second grade (2)	234	1512	432	78	2256	150	2406
Third grade (3)	117	362	1772	205	2456	160	2616
Lowest grade (4)	36	82	179	492	789	50	839
Sub-Total	1907	2222	2507	841	7477	500	7977
Left Only	160	180	200	60	600	*	*
Total	2067	2402	2707	901	8077	*	8577

including Stuart [1955], Bhapkar [1966], Ireland, Ku and Kullback [1969] and GSK. The remaining 1100 observations form artificial data for 600 women for whom only left eye vision was reported and 500 women for whom only right eye vision was reported. It will be presumed that the incomplete data for women with vision classified only for one eye arose in a completely random manner which was statistically independent of the true classification of their vision with respect to both eyes. This assumption allows us to say that the marginal probabilities pertaining to left eye vision and right eye vision for women classified on both eyes are the same parameters as the probabilities pertaining to left eye vision for women classified only for the left eye and to right eye vision for women classified only for the right eye respectively.

To apply the GSK approach to this problem, we view the data as coming from three populations and arrange it in an array as follows:

$$\begin{bmatrix} \underset{\sim}{n}' \\ \underset{\sim}{n}_R' \\ \underset{\sim}{n}_L' \end{bmatrix} = \begin{bmatrix} 1520 & 266 & 124 & 66 & 234 & 1512 & 432 & 78 & 117 & 362 & 1772 & 205 & 36 & 82 & 179 & 492 \\ 140 & 150 & 160 & 50 \\ 160 & 180 & 200 & 60 \end{bmatrix}$$

Let $\pi_{jj'}$ denote the probability that a subject in the population is classified according to the j-th category on the right eye and the j'-th category on the left eye where $j,j'=1,2,3,4$; and let $\pi_{j.}$ and $\pi_{.j'}$ represent corresponding marginal probabilities with

$$\pi_{j.} = \sum_{j'=1}^{4} \pi_{jj'} \quad \text{and} \quad \pi_{.j'} = \sum_{j=1}^{4} \pi_{jj'} \ .$$

If we define $\underset{\sim}{p} = (\underset{\sim}{n}/n)$, $\underset{\sim}{p}_R = (\underset{\sim}{n}_R/n_R)$, and $\underset{\sim}{p}_L = (\underset{\sim}{n}_L/n_L)$ where $n = 7477$, $n_R = 500$, and $n_L = 600$ are the sums of the elements in $\underset{\sim}{n}$, $\underset{\sim}{n}_R$, $\underset{\sim}{n}_L$ respectively and if we let $\underset{\sim}{p}_G' = (\underset{\sim}{p}', \underset{\sim}{p}_R', \underset{\sim}{p}_L')$ be the augmented vector of observed relative frequencies, then

$$\underset{\sim}{E}\{\underset{\sim}{p}_G\} = \underset{\sim}{E} \begin{bmatrix} \underset{\sim}{p} \\ \underset{\sim}{p}_R \\ \underset{\sim}{p}_L \end{bmatrix} = \begin{bmatrix} \underset{\sim}{\pi} \\ \underset{\sim}{\pi}_{*\cdot} \\ \underset{\sim}{\pi}_{\cdot *} \end{bmatrix} \tag{2.1}$$

where

$$\underset{\sim}{\pi}' = (\pi_{11},\pi_{12},\pi_{13},\pi_{14},\pi_{21},\pi_{22},\pi_{23},\pi_{24},\pi_{31},\pi_{32},\pi_{33},\pi_{34},\pi_{41},\pi_{42},\pi_{43},\pi_{44})$$
$$\underset{\sim}{\pi}_{*\cdot}' = (\pi_{1\cdot},\pi_{2\cdot},\pi_{3\cdot},\pi_{4\cdot})$$
$$\underset{\sim}{\pi}_{\cdot *}' = (\pi_{\cdot 1},\pi_{\cdot 2},\pi_{\cdot 3},\pi_{\cdot 4})$$

In addition, the covariance matrix for $\underset{\sim}{p}_G$ is given by

$$\underset{\sim}{V}(\underset{\sim}{\pi}) = \underset{\sim\sim}{Var} \begin{bmatrix} \underset{\sim}{p} \\ \underset{\sim}{p}_R \\ \underset{\sim}{p}_L \end{bmatrix} = \begin{bmatrix} (\underset{\sim}{D}_{\underset{\sim}{\pi}} - \underset{\sim}{\pi}\underset{\sim}{\pi}')/n & \underset{\sim}{0}_{16,4} & \underset{\sim}{0}_{16,4} \\ \underset{\sim}{0}_{4,16} & (\underset{\sim}{D}_{\underset{\sim}{\pi}_{*\cdot}} - \underset{\sim}{\pi}_{*\cdot}\underset{\sim}{\pi}_{*\cdot}')/n_R & \underset{\sim}{0}_{4,4} \\ \underset{\sim}{0}_{4,16} & \underset{\sim}{0}_{4,4} & (\underset{\sim}{D}_{\underset{\sim}{\pi}_{\cdot *}} - \underset{\sim}{\pi}_{\cdot *}\underset{\sim}{\pi}_{\cdot *}')/n_L \end{bmatrix} \tag{2.2}$$

where $\underset{\sim}{0}_{k,k'}$ denotes a (kxk') zero matrix and $\underset{\sim}{D}_{\underset{\sim}{\pi}}$, $\underset{\sim}{D}_{\underset{\sim}{\pi}_{*\cdot}}$, $\underset{\sim}{D}_{\underset{\sim}{\pi}_{\cdot *}}$ are diagonal matrices with diagonal elements being corresponding elements of $\underset{\sim}{\pi}$, $\underset{\sim}{\pi}_{*\cdot}$, $\underset{\sim}{\pi}_{\cdot *}$. The matrix $\underset{\sim}{V}(\underset{\sim}{\pi})$ can be consistently estimated by replacing $\underset{\sim}{\pi}, \underset{\sim}{\pi}_{*\cdot}, \underset{\sim}{\pi}_{\cdot *}$ with $\underset{\sim}{p}$, $\underset{\sim}{p}_R$, $\underset{\sim}{p}_L$ respectively in (2.2); i.e., forming $\underset{\sim}{V}(\underset{\sim}{p}_G)$.

The estimate $\underset{\sim}{p} = (\underset{\sim}{n}/n)$ of $\underset{\sim}{\pi}$ in (2.1) can be adjusted to reflect the available data for all 8577 women by applying weighted least square to the vector $\underset{\sim}{p}_G$ with weights derived from the matrix $\underset{\sim}{V}(\underset{\sim}{p}_G)$ in the manner described by GSK. In order to avoid computational singularities which arise because the sum of the elements in each of the vectors $\underset{\sim}{p}$, $\underset{\sim}{p}_R$, $\underset{\sim}{p}_L$ is unity, the analysis is applied to the vector $\underset{\sim}{F}(\underset{\sim}{p}_G) = \underset{\sim}{A}\underset{\sim}{p}_G$ where

$$\underset{21\times 24}{A} = \begin{bmatrix} \underset{\sim}{I}_{15} & \underset{\sim}{0}_{15,1} & \underset{\sim}{0}_{15,3} & \underset{\sim}{0}_{15,1} & \underset{\sim}{0}_{15,3} & \underset{\sim}{0}_{15,1} \\ \underset{\sim}{0}_{3,15} & \underset{\sim}{0}_{3,1} & \underset{\sim}{I}_3 & \underset{\sim}{0}_{3,1} & \underset{\sim}{0}_{3,3} & \underset{\sim}{0}_{3,1} \\ \underset{\sim}{0}_{3,15} & \underset{\sim}{0}_{3,1} & \underset{\sim}{0}_{3,3} & \underset{\sim}{0}_{3,1} & \underset{\sim}{I}_3 & \underset{\sim}{0}_{3,1} \end{bmatrix}$$

with $\underset{\sim}{I}_3$ and $\underset{\sim}{I}_{15}$ being identity matrices and the other entries being suitable zero matrices. For the vector $\underset{\sim}{F}$, the corresponding weights are derived from $\underset{\sim}{V}_{\underset{\sim}{F}} = \underset{\sim}{A}[\underset{\sim}{V}(\underset{\sim}{p}_G)]\underset{\sim}{A}'$. From (2.1), it follows that the linear model which pertains to $\underset{\sim}{F}$ is given by $E\{\underset{\sim}{F}\} = \underset{\sim}{X}\underset{\sim}{\beta}$ where $\underset{\sim}{X}$ is the known coefficient

$$\underset{21\times 1}{E\{F\}} = \underset{21\times 15}{X}\ \underset{15\times 1}{\beta} = \begin{bmatrix} & & & & & & & & & & & & & & \\ & & & & & & & \underset{\sim}{I}_{15} & & & & & & & \\ & & & & & & & & & & & & & & \\ 1 & 1 & 1 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 1 & 1 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 1 & 1 & 1 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 1 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 1 \end{bmatrix} \underset{\sim}{\beta}, \qquad (2.3)$$

matrix specified in (2.3) and $\underset{\sim}{\beta}$ is the vector of the first 15 components of $\underset{\sim}{\pi}$; i.e. $\underset{\sim}{\pi}$ with π_{44} deleted. The resulting estimate $\underset{\sim}{b}$ of $\underset{\sim}{\beta}$ is given by

$$\underset{\sim}{b} = (\underset{\sim}{X}'\underset{\sim}{V}_{\underset{\sim}{F}}^{-1}\underset{\sim}{X})^{-1}\underset{\sim}{X}'\underset{\sim}{V}_{\underset{\sim}{F}}^{-1}\underset{\sim}{F} \qquad (2.4)$$

with estimated covariance matrix

$$\underset{\sim}{V}_{\underset{\sim}{b}} = \mathrm{Var}(\underset{\sim}{b}) = (\underset{\sim}{X}'\underset{\sim}{V}_{\underset{\sim}{F}}^{-1}\underset{\sim}{X})^{-1} \qquad (2.5)$$

A goodness of fit statistic for assessing the extent to which the model (2.3) characterizes the data is

$$X^2 = SS(\underset{\sim}{E}\{\underset{\sim}{F}\} = \underset{\sim\sim}{X}\underset{\sim}{\beta}) = \underset{\sim}{F}'\underset{\sim\sim}{V}_{\underset{\sim}{F}}^{-1}\underset{\sim}{F} - \underset{\sim}{b}'(\underset{\sim}{X}'\underset{\sim\sim}{V}_{\underset{\sim}{F}}^{-1}\underset{\sim}{X})\underset{\sim}{b} \qquad (2.6)$$

which has approximately a chi-square distribution with D.F. = {(No. of rows in $\underset{\sim}{X}$) - (No. of columns in $\underset{\sim}{X}$)} in large samples, under the hypothesis that the model fits. For these data, X^2=2.33 with D.F. = 6 which is non-significant (α=.25), and thus further consideration of this model is justified. Finally, it can be argued that the test statistic (2.6) provides a means of partially checking the validity of the fundamental assumption that the incomplete data arose in a completely random manner. Caution, however, should be exercised in such interpretations since the 6 degrees of freedom in (2.6) here apply only to comparisons of corresponding marginal distributions in the three samples.

Since the model (2.3) adequately describes the data, tests of hypotheses with respect to the parameters comprising $\underset{\sim}{\beta}$ can be undertaken. In particular, for a general hypothesis of the form $H_0 : \underset{\sim\sim}{C}\underset{\sim}{\beta} = \underset{\sim}{0}$ where $\underset{\sim}{C}$ is a known (d x 15) matrix of full rank, a suitable test statistic is

$$X^2 = SS(\underset{\sim\sim}{C}\underset{\sim}{\beta} = \underset{\sim}{0}) = \underset{\sim}{b}'\underset{\sim}{C}'[\underset{\sim}{C}(\underset{\sim}{X}'\underset{\sim\sim}{V}_{\underset{\sim}{F}}^{-1}\underset{\sim}{X})^{-1}\underset{\sim}{C}']^{-1}\underset{\sim\sim}{C}\underset{\sim}{b} \qquad (2.7)$$

which has approximately a chi-square distribution with D.F. = d in large samples under H_0. An hypothesis which is of particular interest for these data is marginal homogeneity (marginal symmetry). This hypothesis may be written as

$$H_0 : \pi_{j.} = \pi_{.j} \text{ for } j = 1,2,3,4 \qquad (2.8)$$

which in terms of $\underset{\sim}{\beta}$ corresponds to

$$\underset{\sim\sim}{C}\underset{\sim}{\beta} = \begin{bmatrix} 0 & 1 & 1 & 1 & -1 & 0 & 0 & 0 & -1 & 0 & 0 & 0 & -1 & 0 & 0 & 0 \\ 0 & -1 & 0 & 0 & 1 & 0 & 1 & 1 & 0 & -1 & 0 & 0 & 0 & -1 & 0 & 0 \\ 0 & 0 & -1 & 0 & 0 & 0 & -1 & 0 & 1 & 1 & 0 & 1 & 0 & 0 & -1 & 0 \end{bmatrix} \underset{\sim}{\beta} = \underset{\sim}{0}. \qquad (2.9)$$

For $\underset{\sim}{C}$ as specified in (2.9), the statistic (2.7) is $X^2 = 11.41$ with D.F. = 3, which is statistically significant (α=.01). When this hypothesis was investigated by GSK for the original data, ignoring the artificial data for the 1100 incompletely classified women, $X^2 = 11.98$ was obtained. In some sense this result is surprising since one would usually expect X^2-statistics of this type to increase as more data were used in the analysis. However, here the additional data caused the differences between the respective estimates of the marginal probabilities for right eye and left eye to become smaller (see Table 2) than they were originally. On the other hand, Table 2 shows that the estimated

TABLE 2

ESTIMATED PROBABILITIES & STANDARD ERRORS FOR EYE DATA

Cell	Data for Completely Classified Women (n=7477)		Data For All Women (n=8577)	
	p	s.e. (p)	$\hat{\pi}$	s.e. ($\hat{\pi}$)
(1,1)	.20329	.00465	.20456	.00443
(1,2)	.03557	.00214	.03562	.00213
(1,3)	.01658	.00148	.01654	.00147
(1,4)	.00882	.00108	.00876	.00108
(2,1)	.03129	.00201	.03146	.00200
(2,2)	.20222	.00464	.20231	.00446
(2,3)	.05777	.00270	.05759	.00267
(2,4)	.01043	.00118	.01035	.00117
(3,1)	.01564	.00144	.01576	.00143
(3,2)	.04841	.00248	.04851	.00246
(3,3)	.23699	.00492	.23664	.00471
(3,4)	.02741	.00189	.02725	.00187
(4,1)	.00481	.00080	.00482	.00080
(4,2)	.01096	.00120	.01093	.00120
(4,3)	.02394	.00177	.02378	.00175
(4,4)	.06580	.00287	.06507	.00275
(1,.)	.26428	.00510	.26549	.00484
(2,.)	.30173	.00530	.30172	.00506
(3,.)	.32847	.00543	.32817	.00517
(4,.)	.10552	.00355	.10462	.00338
(.,1)	.25505	.00504	.25661	.00480
(.,2)	.29718	.00529	.29738	.00507
(.,3)	.33529	.00546	.33456	.00522
(.,4)	.11248	.00365	.11144	.00349

standard errors for the estimates $\hat{\underset{\sim}{\pi}}$ of the cell probabilities derived using all the data are uniformly smaller than the estimated standard errors for $\underset{\sim}{p}$ which reflect only data for completely classified women; a similar statement applies to estimates of right eye and left eye marginal probabilities derived from $\hat{\underset{\sim}{\pi}}$ and $\underset{\sim}{p}$ respectively. Thus, by including the incompletely classified women in the analysis, more precise estimators of these parameters have been obtained in the sense of estimated standard error.

The above analysis has emphasized the use of the minimum-χ^2_1 estimates $\hat{\underset{\sim}{\pi}}$. Alternatively, Reinfurt [1970] has implied that if the analysis here is repeated but with $\underset{\sim}{V}(\underset{\sim}{\pi})$ estimated by $V(\hat{\underset{\sim}{\pi}})$ rather than $\underset{\sim}{V}(\underset{\sim}{p}_G)$ and the process is continued in an iterative manner until successive estimates of $\underset{\sim}{\pi}$ are satisfactorily similar, the maximum likelihood estimate is obtained. However, this result as well as that of Hocking and Oxspring [1971] only applies to the unrestricted model (i.e., exclusive of any hypotheses on the underlying parameters). Maximum likelihood estimation of $\underset{\sim}{\pi}$ under hypotheses like (2.8) poses more difficult mathematical problems for situations where there is additional and incomplete categorical data of this type.

2.2 An incomplete split-plot experiment

Let us next consider a hypothetical experiment which has been undertaken to compare three drugs A,B, and C. Suppose the response of a given subject to any of these drugs can be observed in a quantal sense as either favorable or unfavorable. For $n_{ABC} = 46$ subjects, each of the three drugs was administered and the separate responses to each were noted; for $n_{AB} = 28$ subjects, only A and B were administered; for $n_A = 16$ subjects, only A. Groups of $n_{AC} = 25$, $n_{BC} = 26$, $n_B = 15$, $n_C = 14$ subjects were treated analogously. The observed results are given in Table 3.

TABLE 3

RESPONSES TO DRUGS A,B,C
(1 DENOTES FAVORABLE RESPONSE, 0 DENOTES UNFAVORABLE RESPONSE, AND * DENOTES THAT THE DRUG WAS NOT RECEIVED)

PATTERNS OF RESPONSE			NUMBER OF RESPONDENTS	
A	B	C	TOTAL = 170	
1	1	1	6	
1	1	0	16	
1	0	1	2	
1	0	0	4	
0	1	1	2	
0	1	0	4	
0	0	1	6	
0	0	0	6	$n_{ABC} = 46$
1	1	*	12	
1	0	*	4	
0	1	*	4	
0	0	*	8	$n_{AB} = 28$
1	*	1	5	
1	*	0	10	
0	*	1	4	
0	*	0	6	$n_{AC} = 25$
*	1	1	4	
*	1	0	12	
*	0	1	5	
*	0	0	5	$n_{BC} = 26$
1	*	*	10	
0	*	*	6	$n_{A} = 16$
*	1	*	11	
*	0	*	4	$n_{B} = 15$
*	*	1	5	
*	*	0	9	$n_{C} = 14$
				TOTAL = 170

For the 46 women who received all three drugs, the frequencies of the respective responses come from an illustrative example that has been treated by several authors including Cochran [1950], Bhapkar [1965], and GSK. The data for the other 124 subjects are artificial.

In this example, the incomplete nature of the data is considered to have been built into the experimental design. Thus, if the set of drugs a subject receives is determined randomly, it is appropriate to assume that the same parameters π_A, π_B, π_C represent the probabilities of a favorable response to drugs A, B, C respectively for each of the groups of subjects.

The data in this example can be analyzed in a manner similar to that demonstrated in Section 2.1. First of all, we view the results in Table 3 as coming from seven populations and arrange it in the array (2.10). Let

$$\begin{bmatrix} \underset{\sim}{n}'_{ABC} \\ \underset{\sim}{n}'_{AB} \\ \underset{\sim}{n}'_{AC} \\ \underset{\sim}{n}'_{BC} \\ \underset{\sim}{n}'_{A} \\ \underset{\sim}{n}'_{B} \\ \underset{\sim}{n}'_{C} \end{bmatrix} = \begin{bmatrix} 6 & 16 & 2 & 4 & 2 & 4 & 6 & 6 \\ 12 & 4 & 4 & 8 & & & & \\ 5 & 10 & 4 & 6 & & & & \\ 4 & 12 & 5 & 5 & & & & \\ 10 & 6 & & & & & & \\ 11 & 4 & & & & & & \\ 5 & 9 & & & & & & \end{bmatrix} \qquad (2.10)$$

$\underset{\sim}{p}_i = (\underset{\sim}{n}_i / n_i)$ where i takes on the values i = ABC, AB, AC, BC, A, B, C, and let $\underset{\sim}{p}_G$ be the compound vector $\underset{\sim}{p}'_G = (\underset{\sim}{p}'_{ABC}, \underset{\sim}{p}'_{AB}, \underset{\sim}{p}'_{AC}, \underset{\sim}{p}'_{BC}, \underset{\sim}{p}'_{A}, \underset{\sim}{p}'_{B}, \underset{\sim}{p}'_{C})$. If we then form the vector of linear functions $\underset{\sim}{F} = \underset{\sim\sim}{A}\underset{\sim}{p}_G$ where $\underset{\sim}{A}$ is defined in (2.11),

$$\underset{12\times 26}{A} = \begin{bmatrix} 1&1&1&1&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0 \\ 1&1&0&0&1&1&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0 \\ 1&0&1&0&1&0&1&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0 \\ 0&0&0&0&0&0&0&1&1&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0 \\ 0&0&0&0&0&0&0&1&0&1&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0 \\ 0&0&0&0&0&0&0&0&0&0&1&1&0&0&0&0&0&0&0&0&0&0&0&0&0&0 \\ 0&0&0&0&0&0&0&0&0&0&1&0&1&0&0&0&0&0&0&0&0&0&0&0&0&0 \\ 0&0&0&0&0&0&0&0&0&0&0&0&0&1&1&0&0&0&0&0&0&0&0&0&0&0 \\ 0&0&0&0&0&0&0&0&0&0&0&0&1&0&1&0&0&0&0&0&0&0&0&0&0&0 \\ 0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&1&0&0&0&0&0&0&0 \\ 0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&1&0&0&0&0&0&0 \\ 0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&1&0&0&0&0&0 \end{bmatrix} \qquad (2.11)$$

then it follows that $\underset{\sim}{E}\{\underset{\sim}{F}\} = \underset{\sim\sim}{X}\underset{\sim}{\pi}$ where $\underset{\sim}{X}$ is given in (2.12) and $\underset{\sim}{\pi}' = (\pi_A, \pi_B, \pi_C)$.

$$\underset{\sim}{E\{F\}} = \underset{\sim}{X}\underset{\sim}{\pi} = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \\ 1 & 0 & 0 \\ 0 & 1 & 0 \\ 1 & 0 & 0 \\ 0 & 0 & 1 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \\ 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} \pi_A \\ \pi_B \\ \pi_C \end{bmatrix} \qquad (2.12)$$

A consistent estimate $\underset{\sim}{V}_{\underset{\sim}{F}}$ for the covariance matrix of $\underset{\sim}{F}$ is given by $\underset{\sim}{V}_{\underset{\sim}{F}} = \underset{\sim}{A}[\underset{\sim}{V}(\underset{\sim}{p}_G)]\underset{\sim}{A}'$ with $\underset{\sim}{V}(\underset{\sim}{p}_G)$ being the block diagonal matrix shown in (2.13) which has

$$\underset{26x26}{\underset{\sim}{V}(\underset{\sim}{p}_G)} = \begin{bmatrix} \underset{\sim}{V}_{ABC}(\underset{\sim}{p}_{ABC}) & \underset{\sim}{0}_{8,4} & \underset{\sim}{0}_{8,4} & \underset{\sim}{0}_{8,4} & \underset{\sim}{0}_{8,2} & \underset{\sim}{0}_{8,2} & \underset{\sim}{0}_{8,2} \\ & \underset{\sim}{V}_{AB}(\underset{\sim}{p}_{AB}) & \underset{\sim}{0}_{4,4} & \underset{\sim}{0}_{4,4} & \underset{\sim}{0}_{4,2} & \underset{\sim}{0}_{4,2} & \underset{\sim}{0}_{4,2} \\ & & \underset{\sim}{V}_{AC}(\underset{\sim}{p}_{AC}) & \underset{\sim}{0}_{4,4} & \underset{\sim}{0}_{4,2} & \underset{\sim}{0}_{4,2} & \underset{\sim}{0}_{4,2} \\ & & & \underset{\sim}{V}_{BC}(\underset{\sim}{p}_{BC}) & \underset{\sim}{0}_{4,2} & \underset{\sim}{0}_{4,2} & \underset{\sim}{0}_{4,2} \\ & & & & \underset{\sim}{V}_{A}(\underset{\sim}{p}_{A}) & \underset{\sim}{0}_{2,2} & \underset{\sim}{0}_{2,2} \\ & & & & & \underset{\sim}{V}_{B}(\underset{\sim}{p}_{B}) & \underset{\sim}{0}_{2,2} \\ & & & & & & \underset{\sim}{V}_{C}(\underset{\sim}{p}_{C}) \end{bmatrix} \qquad (2.13)$$

sub-matrices $\underset{\sim}{V}_i(\underset{\sim}{p}_i) = (\underset{\sim}{D}_{\underset{\sim}{p}_i} - \underset{\sim}{p}_i\underset{\sim}{p}_i')/n_i$ for i = ABC, AB, AC, BC, A, B, C, respectively on the main diagonal. Following the approach discussed in the preceding section, the model (2.12) is fitted to the vector $\underset{\sim}{F}$ by weighted least squares. The resulting estimates for π_A, π_B, π_C and their estimated standard errors as determined from

expressions analogous to (2.4) and (2.5) are given in the last column of Table 4.

TABLE 4

ESTIMATED PROBABILITIES AND (STANDARD ERRORS)

PARAMETERS	SUBJECTS RECEIVING ALL THREE DRUGS (n=46)	SUBJECTS RECEIVING ONLY ONE OR TWO DRUGS (n=124)	ALL SUBJECTS (n=170)
π_A	.609 (.072)	.604 (.058)	.607 (.044)
π_B	.609 (.072)	.629 (.056)	.621 (.044)
π_C	.348 (.070)	.352 (.059)	.350 (.045)
X^2-statistic (D.F. = 2) $H_0: \pi_A=\pi_B=\pi_C$	6.58	12.03	18.79

A test of the hypothesis H_0: $\pi_A = \pi_B = \pi_C$ may be performed by using

$$\underset{\sim}{C} = \begin{bmatrix} 1 & -1 & 0 \\ 1 & 0 & -1 \end{bmatrix}$$

in an expression analogous to (2.7). The resulting $X^2 = 18.79$ with D.F. = 2 is statistically significant (α=.01). Finally, it is worthwhile to note that the goodness of fit statistic analogous to (2.6) for the model (2.12) is $X^2 = 1.37$ with D.F. = 9; thus the assumption that the same parameters π_A, π_B, π_C represent the probabilities of a favorable response to drugs A, B, C respectively for each of the groups of subjects where they apply is consistent with the data.

For comparative purposes, results are also separately shown in Table 4 for the subjects receiving all three drugs and for the subjects receiving only one or two drugs. These were obtained by applying the previous type of analysis to the corresponding rows of (2.12); i.e., the first three rows for the former

and the last nine rows for the latter. For the hypothesis $H_0 : \pi_A = \pi_B = \pi_C$ $X^2 = 6.58$ with D.F. = 2 and $X^2 = 12.03$ with D.F. = 2 respectively; and for goodness of fit, there is no test for the former and $X^2 = 1.29$ with D.F. = 6 for the latter.

2.3 An application with supplemental samples

This example is based on the data considered by Dyke and Patterson [1952], Bishop [1969], Goodman [1970], and other authors. Suppose in order to study the relationship between an individual's knowledge of cancer and extent of contact with reading materials, three samples were selected without replacement from a large population by simple random sampling (more complex sampling designs could be treated by the approach described in Johnson and Koch [1970] and Koch and Reinfurt [1970]). In sample I the individuals were cross classified according to the following three categorical variables:

i. whether they read newspapers or not,

ii. whether they read books or magazines (solid reading) or not,

iii. whether their knowledge of cancer was good or poor.

In sample II, subjects were classified only by (i) and (iii); in Sample III, only by (ii) and (iii).

The observed data are given in Table 5. For the 1729 subjects in sample I, the frequencies are the appropriate marginal totals of the data considered by Dyke and Patterson [1952] and other authors. The data for the other 910 subjects are artificial.

The incomplete nature of the data here is considered to have resulted from the structure of the sample survey design. Samples II and III might be separately conducted preliminary investigations involving variables (i) and (ii). Alternatively, in many surveys it can be expected that response errors due to respondent

hostility or other causes will increase with the size of a questionnaire. In such situations it may be advantageous to split the questionnaire, resulting in data such as that of Table 5. In either case, the analytical strategy to be applied is similiar to that of Section 2.2. However, for these data, attention will be directed at the relationship among the three categorical variables as it pertains to the conditional

TABLE 5

RESPONSES TO CANCER SURVEYS
(* DENOTES UNMEASURED RESPONSE)

PATTERNS OF RESPONSE			NUMBER OF RESPONDENTS	
NEWSPAPERS	SOLID READING	CANCER KNOWLEDGE		TOTAL = 2639
YES	YES	GOOD	353	
YES	YES	POOR	270	
YES	NO	GOOD	125	
YES	NO	POOR	225	
NO	YES	GOOD	87	
NO	YES	POOR	110	
NO	NO	GOOD	103	
NO	NO	POOR	456	$n_I = 1729$
YES	*	GOOD	90	
YES	*	POOR	100	
NO	*	GOOD	40	
NO	*	POOR	110	$n_{II} = 340$
*	YES	GOOD	150	
*	YES	POOR	120	
*	NO	GOOD	80	
*	NO	POOR	220	$n_{III} = 570$

probability of good knowledge of cancer given fixed categories for exposure to newspapers and solid reading.

Following the approach described in Sections 2.1 and 2.2, we first view the results in Table 5 as coming from three populations and arrange it in an array as follows:

$$\begin{bmatrix} \underset{\sim}{n}'_I \\ \underset{\sim}{n}'_{II} \\ \underset{\sim}{n}'_{III} \end{bmatrix} = \begin{bmatrix} 353 & 270 & 125 & 225 & 87 & 110 & 103 & 456 \\ 90 & 100 & 40 & 110 & & & & \\ 150 & 120 & 80 & 220 & & & & \end{bmatrix}$$

Let $\underset{\sim}{p}_I = (\underset{\sim}{n}_I/n_I)$, $\underset{\sim}{p}_{II} = (\underset{\sim}{n}_{II}/n_{II})$, and $\underset{\sim}{p}_{III} = (\underset{\sim}{n}_{III}/n_{III})$ where $n_I = 1729$, $n_{II} = 340$, $n_{III} = 570$ are the sizes of the three samples. Let $\underset{\sim}{p}_G$ be the augmented vector $\underset{\sim}{p}'_G = (\underset{\sim}{p}'_I, \underset{\sim}{p}'_{II}, \underset{\sim}{p}'_{III})$. If we then form the vector of linear functions $\underset{\sim}{F} = \underset{\sim\sim}{A}\underset{\sim}{p}_G$ where

$$\underset{13\times16}{\underset{\sim}{A}} = \begin{bmatrix} \underset{\sim}{I}_7 & \underset{\sim}{0}_{7,1} & \underset{\sim}{0}_{7,3} & \underset{\sim}{0}_{7,1} & \underset{\sim}{0}_{7,3} & \underset{\sim}{0}_{7,1} \\ \underset{\sim}{0}_{3,7} & \underset{\sim}{0}_{3,1} & \underset{\sim}{I}_3 & \underset{\sim}{0}_{3,1} & \underset{\sim}{0}_{3,3} & \underset{\sim}{0}_{3,1} \\ \underset{\sim}{0}_{3,7} & \underset{\sim}{0}_{3,1} & \underset{\sim}{0}_{3,3} & \underset{\sim}{0}_{3,1} & \underset{\sim}{I}_3 & \underset{\sim}{0}_{3,1} \end{bmatrix}$$

then it follows that $\underset{\sim}{E}\{\underset{\sim}{F}\} = \underset{\sim\sim}{X}\underset{\sim}{\pi}_T$, where $\underset{\sim}{X}$ and $\underset{\sim}{\pi}_T$ are given in (2.14). The

$$\underset{\sim}{E}\{\underset{\sim}{F}\} = \underset{\sim\sim}{X}\underset{\sim}{\pi}_T = \begin{bmatrix} 1 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 1 \\ 1 & 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 & 1 \\ 1 & 0 & 0 & 0 & 1 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} \pi_{111} \\ \pi_{112} \\ \pi_{121} \\ \pi_{122} \\ \pi_{211} \\ \pi_{212} \\ \pi_{221} \end{bmatrix} \qquad (2.14)$$

elements $\pi_{hh'j}$ which comprise $\underset{\sim}{\pi}_T$ may be interpreted as the probabilities that an individual is classified into the h-th category with respect to newspapers (1 if yes and 2 if no), the h'-th category with respect to solid reading (1 if yes and 2 if no), and the j-th category with respect to knowledge of cancer (1 if good and

2 if poor). The estimated covariance matrix $\underset{\sim}{V}_F$ for $\underset{\sim}{F}$ is given by $\underset{\sim}{V}_F = \underset{\sim}{A}[\underset{\sim}{V}(\underset{\sim}{p}_G)]\underset{\sim}{A}'$, $\underset{\sim}{V}(\underset{\sim}{p}_G)$ being the following block diagonal matrix

$$\underset{\sim}{V}(\underset{\sim}{p}_G) = \begin{bmatrix} \underset{\sim}{V}_I(\underset{\sim}{p}_I) & \underset{\sim}{0}_{7,3} & \underset{\sim}{0}_{7,3} \\ & \underset{\sim}{V}_{II}(\underset{\sim}{p}_{II}) & \underset{\sim}{0}_{3,3} \\ & & \underset{\sim}{V}_{III}(\underset{\sim}{p}_{III}) \end{bmatrix}$$

where $\underset{\sim}{V}(\underset{\sim}{p}_i) = (\underset{\sim}{D}_{\underset{\sim}{p}_i} - \underset{\sim}{p}_i\underset{\sim}{p}_i')/n_i$ for i = I, II, III.

The vector $\underset{\sim}{\pi}_T$ is estimated by applying weighted least squares to the model (2.14). The goodness of fit statistic analogous to (2.6) for the model (2.14) is $X^2 = 1.02$ with D.F. = 6. Hence, the model is consistent with the data. The resulting estimates $\hat{\underset{\sim}{\pi}}_T$ and their standard errors are shown in the last two columns of Table 6. In addition, $\hat{\pi}_{222}=(1 - \underset{\sim}{j}_7'\hat{\underset{\sim}{\pi}}_T)$ and its standard error, where $\underset{\sim}{j}_7$ is a vector of ones, are included here. Finally, corresponding estimates from the Sample I data only are given for purposes of comparison.

TABLE 6

ESTIMATED PROBABILITIES FOR DYKE-PATTERSON DATA

PATTERN OF RESPONSE			SAMPLE I		ALL SAMPLES	
NEWSPAPERS	SOLID READING	CANCER KNOWLEDGE	p	s.e.p.	$\hat{\pi}$	s.e.$\hat{\pi}$
Y	Y	G	.20416	.00969	.20433	.00833
Y	Y	P	.15616	.00873	.15529	.00770
Y	N	G	.07230	.00623	.07278	.00573
Y	N	P	.13013	.00809	.13024	.00760
N	Y	G	.05032	.00526	.05127	.00498
N	Y	P	.06362	.00587	.06284	.00561
N	N	G	.05957	.00569	.06105	.00520
N	N	P	.26374	.01060	.26220	.00929

In order to estimate the conditional probabilities of good knowledge of cancer for fixed categories with respect to newspapers and solid reading, several additional calculations are required. If we let $\hat{\underset{\sim}{\pi}}$ be defined by $\hat{\underset{\sim}{\pi}}' = (\hat{\underset{\sim}{\pi}}_T', \hat{\pi}_{222})$, then a consistent estimate for the covariance matrix for $\hat{\underset{\sim}{\pi}}$ is

$$\hat{\underset{\sim}{V}} = \begin{bmatrix} \underset{\sim}{I}_7 \\ -\underset{\sim}{j}_7' \end{bmatrix} [\underset{\sim}{X}' \underset{\sim}{V}_F^{-1} \underset{\sim}{X}]^{-1} [\underset{\sim}{I}_7, -\underset{\sim}{j}_7] \qquad (2.15)$$

Let $f_{hh'} = \{\hat{\pi}_{hh'1}/(\hat{\pi}_{hh'1} + \hat{\pi}_{hh'2})\}$ denote the estimated conditional probability for good knowledge of cancer for individuals in the h-th category with respect to newspapers and the h'-th category with respect to solid reading. Let $\underset{\sim}{f}$ be the vector defined by $\underset{\sim}{f}' = (f_{11}, f_{12}, f_{21}, f_{22})$. The vector $\underset{\sim}{f}$ may be written in matrix notation as $\underset{\sim}{f} = \underset{\sim\sim}{\exp}\{\underset{\sim}{K}[\underset{\sim\sim}{\log}_e(\underset{\sim}{A}_f \hat{\underset{\sim}{\pi}})]\}$ where $\underset{\sim}{A}_f$ and $\underset{\sim}{K}$ are displayed in (2.16);

$$\underset{\sim}{A}_f = \begin{bmatrix} 1&0&0&0&0&0&0&0 \\ 1&1&0&0&0&0&0&0 \\ 0&0&1&0&0&0&0&0 \\ 0&0&1&1&0&0&0&0 \\ 0&0&0&0&1&0&0&0 \\ 0&0&0&0&1&1&0&0 \\ 0&0&0&0&0&0&1&0 \\ 0&0&0&0&0&0&1&1 \end{bmatrix} \quad \text{and} \quad \underset{\sim}{K} = \begin{bmatrix} 1&-1&0&0&0&0&0&0 \\ 0&0&1&-1&0&0&0&0 \\ 0&0&0&0&1&-1&0&0 \\ 0&0&0&0&0&0&1&-1 \end{bmatrix} \qquad (2.16)$$

also, $\underset{\sim\sim}{\log}_e$ transforms a vector to the corresponding vector of logarithms and $\underset{\sim\sim}{\exp}$ transforms a vector to the corresponding vector of anti-logarithms (i.e., exponential functions). A consistent estimate for the covariance matrix of $\underset{\sim}{f}$ is

$$\underset{\sim}{V}_f = \underset{\sim}{D}_f \underset{\sim}{K} \underset{\sim}{D}_g^{-1} \underset{\sim}{A}_f \hat{\underset{\sim}{V}} \underset{\sim}{A}_f' \underset{\sim}{D}_g^{-1} \underset{\sim}{K}' \underset{\sim}{D}_f$$

where $\underset{\sim}{D}_f$ and $\underset{\sim}{D}_g$ are diagonal matrices formed from the vectors $\underset{\sim}{f}$ and $\underset{\sim}{g} = \underset{\sim}{A}_f \hat{\underset{\sim}{\pi}}$ respectively and $\hat{\underset{\sim}{V}}$ is the matrix in (2.15).

On applying these results to the data in this example, we obtain

$$\underset{\sim}{f} = \begin{bmatrix} .568 \\ .358 \\ .449 \\ .189 \end{bmatrix} \quad \text{and} \quad \underset{\sim}{V}_{\underset{\sim}{f}} = \begin{bmatrix} 2.937 & -.289 & -.651 & .039 \\ & 5.485 & .133 & -.424 \\ & & 11.120 & -.348 \\ & & & 2.172 \end{bmatrix} \times 10^{-4}$$

In order to determine the influence that newspapers and solid readings have on the probability of good knowledge of cancer, we fit the model

$$\underset{\sim}{E}\{\underset{\sim}{f}\} = \underset{\sim}{X}_{\underset{\sim}{f}}\underset{\sim}{\beta} = \begin{bmatrix} 1 & 1 & 1 \\ 1 & 1 & -1 \\ 1 & -1 & 1 \\ 1 & -1 & -1 \end{bmatrix} \cdot \begin{bmatrix} \beta_1 \\ \beta_2 \\ \beta_3 \end{bmatrix} \qquad (2.17)$$

by weighted least squares where β_1 is an overall mean, β_2 is a newspaper effect, and β_3 is a solid reading effect. The resulting estimate $\underset{\sim}{b}$ of $\underset{\sim}{\beta}$ and its estimated covariance matrix $\underset{\sim}{V}_{\underset{\sim}{b}}$ are

$$\underset{\sim}{b} = \begin{bmatrix} .385 \\ .076 \\ .115 \end{bmatrix} \quad \text{and} \quad \underset{\sim}{V}_{\underset{\sim}{b}} = \begin{bmatrix} .830 & -.101 & .228 \\ & 1.275 & -.623 \\ & & 1.220 \end{bmatrix} \times 10^{-4}$$

The goodness of fit test analogous to (2.6) for the model (2.15) is $X^2 = 1.01$ with D.F.=1. Thus, the model is supported by the data, which implies that there is no interaction between the effects of newspapers and solid reading with respect to the vector $\underset{\sim}{f}$. Finally, for the hypothesis $H_0 : \beta_2 = 0$, $X^2 = 44.74$ with D.F.=1; and for the hypothesis $H_0 : \beta_3 = 0$, $X^2 = 108.50$ with D.F.=1. Thus, both of these main effects are statistically significant ($\alpha = .01$).

If only the data in sample I are considered, the results corresponding to $\underset{\sim}{b}$ and $\underset{\sim}{V}_{b}$ are

$$\underset{\sim}{b}_{I} = \begin{bmatrix} .382 \\ .078 \\ .115 \end{bmatrix} \quad \text{and} \quad \underset{\sim}{V}_{b_I} = \begin{bmatrix} 1.230 & -.076 & .308 \\ & 1.480 & -.695 \\ & & 1.553 \end{bmatrix} \times 10^{4}$$

Similarly, the goodness of fit test statistic as well as those for H_0: $\beta_2 = 0$ and H_0: $\beta_3 = 0$ are $X^2 = 0.89$, $X^2 = 40.92$ and $X^2 = 84.45$, respectively, each with D.F. = 1. Thus, it is apparent that inclusion of Samples II and III in the analysis produced some improvement in the precision of the estimate of $\underset{\sim}{\beta}$.

2.4 A Graeco-Latin square split-plot experiment

The data in Table 7 represent hypothetical results of a typical experiment conducted on three specially instrumented cars by the University of North Carolina Highway Safety Research Center. The objective here is to determine the effect of three different driver training methods and three different examining procedures on an individual's ability to pass a driving skill test. Representative subjects from populations corresponding to the three driver training methods participate in a driving skill test at three different times under three different combinations of car and examining procedure. These combinations are determined according to the Graeco-Latin square in Table 8, where rows represent the driver training method; columns, the three trials; Latin letters, the three cars; and Greek letters, the three testing procedures. In some cases, certain subjects do

TABLE 7

RESULTS OF DRIVING TESTS
(P = PASS, F = FAIL, * = NOT TAKEN)

PATTERNS OF RESPONSE TRIAL			NUMBER OF RESPONDENTS GROUP		
I	II	III	A	B	C
P	P	P	15	12	6
P	P	F	24	3	14
P	F	P	9	22	6
P	F	F	15	9	6
F	P	P	5	16	10
F	P	F	13	13	25
F	F	P	10	13	13
F	F	F	9	12	20
			$n_{A1} = n_{B1} = n_{C1} = 100$		
P	*	P	6	14	3
P	*	F	21	8	6
F	*	P	8	10	11
F	*	F	5	8	20
			$n_{A2} = n_{B2} = n_{C2} = 40$		
P	P	*	18	8	5
P	F	*	10	11	5
F	P	*	7	8	18
F	F	*	5	13	12
			$n_{A3} = n_{B3} = n_{C3} = 40$		

TABLE 8

DESIGN OF INSTRUMENTED CAR EXPERIMENT

	1	2	3
A	$X\alpha$	$Y\beta$	$Z\gamma$
B	$Z\beta$	$X\gamma$	$Y\alpha$
C	$Y\gamma$	$Z\alpha$	$X\beta$

not participate in all three tests either because of time scheduling problems or because of cost considerations. Thus, the data in Table 7 includes groups of subjects with incomplete response vectors which may be viewed as having arisen either in the context of the missing data in Section 2.1 or the supplemental sample data in Section 2.3.

As in the other examples, we first view the results in Table 7 as coming from nine populations and thus formulate the array displayed in (2.18).

$$\begin{bmatrix} \underset{\sim}{n}'_{A,1} \\ \underset{\sim}{n}'_{B,1} \\ \underset{\sim}{n}'_{C,1} \\ \underset{\sim}{n}'_{A,2} \\ \underset{\sim}{n}'_{B,2} \\ \underset{\sim}{n}'_{C,2} \\ \underset{\sim}{n}'_{A,3} \\ \underset{\sim}{n}'_{B,3} \\ \underset{\sim}{n}'_{C,3} \end{bmatrix} = \begin{bmatrix} 15 & 24 & 9 & 15 & 5 & 13 & 10 & 9 \\ 12 & 3 & 22 & 9 & 16 & 13 & 13 & 12 \\ 6 & 14 & 6 & 6 & 10 & 25 & 13 & 20 \\ 6 & 21 & 8 & 5 & & & & \\ 14 & 8 & 10 & 8 & & & & \\ 3 & 6 & 11 & 20 & & & & \\ 18 & 10 & 7 & 5 & & & & \\ 8 & 11 & 8 & 13 & & & & \\ 5 & 5 & 18 & 12 & & & & \end{bmatrix} \qquad (2.18)$$

For each $\underset{\sim}{n}_i$ in (2.18) define $\underset{\sim}{p}_i = (\underset{\sim}{n}_i / n_i)$ where i takes on values A1, B1,...,C3 and $n_{A,1} = n_{B,1} = n_{C,1} = 100$, $n_{A,2} = n_{B,2} = n_{C,2} = n_{A,3} = n_{B,3} = n_{C,3} = 40$ are the sums of the elements in $\underset{\sim}{n}_{A,1}$, $\underset{\sim}{n}_{B,1}$, $\underset{\sim}{n}_{C,1}$, $\underset{\sim}{n}_{A,2}$, $\underset{\sim}{n}_{B,2}$, $\underset{\sim}{n}_{C,2}$, $\underset{\sim}{n}_{A,3}$, $\underset{\sim}{n}_{B,3}$, $\underset{\sim}{n}_{C,3}$ respectively. Let $\underset{\sim}{p}_G$ be the compound vector $\underset{\sim}{p}'_G = (\underset{\sim}{p}'_{A,1}, \underset{\sim}{p}'_{B,1}, \underset{\sim}{p}'_{C,1}, \underset{\sim}{p}'_{A,2}, \underset{\sim}{p}'_{B,2}, \underset{\sim}{p}'_{C,2}, \underset{\sim}{p}'_{A,3}, \underset{\sim}{p}'_{B,3}, \underset{\sim}{p}'_{C,3})$. If we form the vector of linear functions $\underset{\sim}{F} = \underset{\sim}{A}\underset{\sim}{p}_G$ where $\underset{\sim}{A}$ is defined in (2.19), then the resulting elements of $\underset{\sim}{F}$ represent estimated probabilities with which subjects in the various groups pass the driving test

on respective trials. Under the assumption that there is no second or higher order interaction among the four factors training method, trial, car, and examining procedure with respect to these probabilities of passing, then it follows that $\underset{\sim}{E}\{\underset{\sim}{F}\} = \underset{\sim}{X}\underset{\sim}{\beta}$ where $\underset{\sim}{X}$ and $\underset{\sim}{\beta}$ are displayed in (2.20). The component

$$\underset{\sim}{A} = \begin{bmatrix}
1111000000000000000000000000000000000000000000000000 \\
1100110000000000000000000000000000000000000000000000 \\
1010101000000000000000000000000000000000000000000000 \\
0000000011110000000000000000000000000000000000000000 \\
0000000011001100000000000000000000000000000000000000 \\
0000000010101010000000000000000000000000000000000000 \\
0000000000000000111100000000000000000000000000000000 \\
0000000000000000110011000000000000000000000000000000 \\
0000000000000000101010100000000000000000000000000000 \\
0000000000000000000000001100000000000000000000000000 \\
0000000000000000000000001010000000000000000000000000 \\
0000000000000000000000000011000000000000000000000000 \\
0000000000000000000000000010100000000000000000000000 \\
0000000000000000000000000001100000000000000000000000 \\
0000000000000000000000000000101000000000000000000000 \\
0000000000000000000000000000011000000000000000000000 \\
0000000000000000000000000000010100000000000000000000 \\
0000000000000000000000000000001100000000000000000000 \\
0000000000000000000000000000000101000000000000000000 \\
0000000000000000000000000000000011000000000000000000 \\
0000000000000000000000000000000010100000000000000000 \\
0000000000000000000000000000000001100000000000000000 \\
0000000000000000000000000000000001010000000000000000
\end{bmatrix} \qquad (2.19)$$

$$\underset{\sim}{E\{F\}} = \underset{\sim}{X\beta} = \begin{bmatrix} 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 \\ 1 & 1 & 1 & 0 & -1 & 0 & -1 & 0 & -1 \\ 1 & 1 & 1 & -1 & 0 & -1 & 0 & -1 & 0 \\ 1 & 0 & -1 & 1 & 1 & -1 & 0 & 0 & -1 \\ 1 & 0 & -1 & 0 & -1 & 1 & 1 & -1 & 0 \\ 1 & 0 & -1 & -1 & 0 & 0 & -1 & 1 & 1 \\ 1 & -1 & 0 & 1 & 1 & 0 & -1 & -1 & 0 \\ 1 & -1 & 0 & 0 & -1 & -1 & 0 & 1 & 1 \\ 1 & -1 & 0 & -1 & 0 & 1 & 1 & 0 & -1 \\ 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 \\ 1 & 1 & 1 & -1 & 0 & -1 & 0 & -1 & 0 \\ 1 & 0 & -1 & 1 & 1 & -1 & 0 & 0 & -1 \\ 1 & 0 & -1 & -1 & 0 & 0 & -1 & 1 & 1 \\ 1 & -1 & 0 & 1 & 1 & 0 & -1 & -1 & 0 \\ 1 & -1 & 0 & -1 & 0 & 1 & 1 & 0 & -1 \\ 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 \\ 1 & 1 & 1 & 0 & -1 & 0 & -1 & 0 & -1 \\ 1 & 0 & -1 & 1 & 1 & -1 & 0 & 0 & -1 \\ 1 & 0 & -1 & 0 & -1 & 1 & 1 & -1 & 0 \\ 1 & -1 & 0 & 1 & 1 & 0 & -1 & -1 & 0 \\ 1 & -1 & 0 & 0 & -1 & -1 & 0 & 1 & 1 \end{bmatrix} \begin{bmatrix} \mu \\ \tau_1 \\ \tau_2 \\ \rho_1 \\ \rho_2 \\ \gamma_1 \\ \gamma_2 \\ \xi_1 \\ \xi_2 \end{bmatrix} \qquad (2.20)$$

elements of $\underset{\sim}{\beta}$ are an overall mean μ, training method effects τ_1 and τ_2, trial effects ρ_1 and ρ_2, car effects γ_1 and γ_2, and examination effects ξ_1 and ξ_2. A consistent estimate for the covariance matrix of $\underset{\sim}{F}$ is $\underset{\sim}{V}_{\underset{\sim}{F}} = \underset{\sim}{A}[\underset{\sim}{V}(\underset{\sim}{p}_G)]\underset{\sim}{A}'$ where $\underset{\sim}{V}(\underset{\sim}{p}_G)$ is a block diagonal matrix with the matrices $\underset{\sim}{V}\{\underset{\sim}{p}_i\} = \{\underset{\sim}{D}_{\underset{\sim}{p}_i} - \underset{\sim}{p}_i\underset{\sim}{p}_i'\}/n_i$ for i = A1, B1, C1, A2, B2, C2, A3, B3, C3 respectively being the main diagonal blocks.

As with the previous examples, the model (2.20)can be fitted to the

vector $\underset{\sim}{F}$ by weighted least squares. The resulting estimates $\underset{\sim}{b}$ of $\underset{\sim}{\beta}$ and their standard errors are given in the last two columns of Table 9. Similarly,

TABLE 9

Parameter	Subjects in All Three Trials (n=300) estimate	s.e.	Subjects in Two Trials Only (n=240) estimate	s.e.	All Subjects (n=540) estimate	s.e.
μ	.482	.016	.481	.023	.482	.013
τ_1	.076	.023	.094	.032	.088	.018
τ_2	-.028	.023	-.022	.033	-.029	.019
ρ_1	.026	.023	.051	.025	.033	.019
ρ_2	-.038	.024	-.053	.034	-.040	.019
γ_1	.016	.024	.003	.033	.011	.019
γ_2	-.024	.021	-.005	.032	-.015	.017
ξ_1	.099	.024	.153	.033	.121	.019
ξ_2	.022	.023	-.014	.032	.009	.019

X^2-tests for the effects of the respective factors and the adequacy of the model are determined as in (2.6) and (2.7). These results are displayed in the last column of Table 10. For purposes of comparison, analogous statistics are also

TABLE 10

Hypothesis	X^2-statistic Subjects in All Three Trials	X^2-statistic Subjects in Two Trials Only	X^2-statistic All Subjects
$\tau_1=\tau_2=0$	11.22	9.95	25.50
$\rho_1=\rho_2=0$	2.56	2.77	4.61
$\gamma_1=\gamma_2=0$	1.34	0.02	0.79
$\xi_1=\xi_2=0$	33.84	25.24	62.52
Residual	0.00	0.58	4.52

separately shown in Tables 9 and 10 for the subjects participating in all three trials and for the subjects participating in only two of the trials. These values were obtained by applying the methods described here to the appropriate rows of (2.20), i.e., the first nine rows for the former case and the last twelve rows for the latter.

All of the results in Table 10 indicate that car effects and trial effects are not important. Hence, a revised model can be fitted to $\underset{\sim}{F}$ with these factors excluded (i.e., columns 4, 5, 6, 7 are deleted from $\underset{\sim}{X}$ and ρ_1, ρ_2, γ_1, γ_2 are deleted from $\underset{\sim}{\beta}$). The corresponding estimates and test statistics for this analysis of the data on all subjects are given in Table 11. Finally, it

TABLE 11

Parameter	Estimate	s.e.	Hypothesis	X^2	D.F.
μ	.480	.013	Residual	9.80	16
τ_1	.087	.018			
τ_2	-.031	.019	$\tau_1 = \tau_2 = 0$	24.81	2
ξ_1	.119	.019			
ξ_2	.007	.018	$\xi_1 = \xi_2 = 0$	61.72	2

is worthwhile noting that if it can be a priori assumed that there are no differences among cars and no differences among trials, then the parameters ρ_1, ρ_2, γ_1, γ_2 correspond to the (training method x examination procedure) interaction effects. In this context, these results imply that the main effects of training method and examination procedure are statistically significant (α=.01), and that the corresponding second order interaction is negligible.

2.5 A growth curve problem

In many areas of research, longitudinal data are collected from subjects at several different points in time. The following example is of interest with respect to the analysis of categorical data arising from such investigations.

Suppose four different diets (I,II,III,IV) designed to reduce blood cholesterol are assigned to four different groups of people. At the end of each of three time periods, a blood sample is taken from each available person and classified as to whether the blood cholesterol is normal or abnormal. For diets II, III, and IV, the subjects were closely supervised (e.g.,within a hospital or clinic environment), and responses for all three time periods were obtained. Subjects on diet I were not as closely supervised, however, and for many of these individuals blood sample data were obtainable only for one or two of the time periods. The resulting artificial data are shown in Table 12. It is assumed that the incomplete data in this situation occurred because of scheduling problems (e.g., subjects in group I lived at home and were not able to attend all three appointments for blood samples either because of other commitments, unrelated illness, travel distance, etc.) and thus may be viewed either in the context of the missing data in Sections 2.1 or 2.4 or the supplemental sample data in Section 2.3.

Let us now arrange the results in Table 12 in the array displayed in (2.21).

$$\begin{bmatrix} \underset{\sim}{n}'_{I,1} \\ \underset{\sim}{n}'_{II,1} \\ \underset{\sim}{n}'_{III,1} \\ \underset{\sim}{n}'_{IV,1} \\ \underset{\sim}{n}'_{I,2} \\ \underset{\sim}{n}'_{I,3} \\ \underset{\sim}{n}'_{I,4} \\ \underset{\sim}{n}'_{I,5} \end{bmatrix} = \begin{bmatrix} 2 & 2 & 8 & 9 & 9 & 15 & 27 & 28 \\ 7 & 2 & 5 & 2 & 31 & 5 & 32 & 6 \\ 16 & 13 & 9 & 3 & 14 & 4 & 15 & 6 \\ 31 & 0 & 6 & 0 & 22 & 2 & 9 & 0 \\ 1 & 4 & 6 & 14 & & & & \\ 1 & 3 & 7 & 14 & & & & \\ 3 & 4 & 8 & 10 & & & & \\ 6 & 19 & & & & & & \end{bmatrix} \qquad (2.21)$$

TABLE 12

RESULTS OF BLOOD CHOLESTEROL TESTS
(N = NORMAL, A = ABNORMAL, * = NOT TAKEN)

PATTERNS OF RESPONSE TIME PERIOD			NUMBER OF RESPONDENTS DIET			
0	1	2	I	II	III	IV
N	N	N	2	7	16	31
N	N	A	2	2	13	0
N	A	N	8	5	9	6
N	A	A	9	2	3	0
A	N	N	9	31	14	22
A	N	A	15	5	4	2
A	A	N	27	32	15	9
A	A	A	28	6	6	0
			$n_{I1} = 100$	$n_{II,1}=90$	$n_{III,1}=80$	$n_{IV,1}=70$
N	N	*	1			
N	A	*	4			
A	N	*	6			
A	A	*	14			
			$n_{I,2}= 25$			
N	*	N	1			
N	*	A	3			
A	*	N	7			
A	*	A	14			
			$n_{I,3}= 25$			
*	N	N	3			
*	N	A	4			
*	A	N	8			
*	A	A	10			
			$n_{I,4}=25$			
N	*	*	6			
A	*	*	19			
			$n_{I,5}=25$			

Let $\underset{\sim}{p}_i = (\underset{\sim}{n}_i/n_i)$ for i = I1, II1, III1, IV1, I2, I3, I4, I5, respectively where n_i denotes the sum of the elements in the vector $\underset{\sim}{n}_i$; i.e., $n_{I,1} = 100$, $n_{II,1} = 90$, $n_{III,1} = 80$, $n_{IV,1} = 70$, $n_{I,2} = n_{I,3} = n_{I,4} = n_{I,5} = 25$. Let $\underset{\sim}{p}_G$ be the compound vector

$$\underset{\sim}{p}'_G = (\underset{\sim}{p}'_{I,1}, \underset{\sim}{p}'_{II,1}, \underset{\sim}{p}'_{III,1}, \underset{\sim}{p}_{IV,1}, \underset{\sim}{p}'_{I,2}, \underset{\sim}{p}'_{I,3}, \underset{\sim}{p}'_{I,4}, \underset{\sim}{p}'_{I,5}) \qquad (2.22)$$

If we form the vector of linear functions $\underset{\sim}{F} = \underset{\sim}{A}\underset{\sim}{p}_G$ where $\underset{\sim}{A}$ is defined in (2.23), then the resulting elements of $\underset{\sim}{F}$ represent the estimated probabilities of normal and abnormal blood cholesterol for the respective time periods in the four groups of subjects. If the trend according to which the probability of normal

$$\underset{\sim}{A}_{38\times46} = \begin{bmatrix}
11110000 & & & & & & & & & & & & & & & & & & & & & & & \\
00001111 & & & & & & & & & & & & & & & & & & & & & & & \\
11001100 & & & & & & & & & & & & & & & & & & & & & & & \\
00110011 & & & & & & & & & & & & & & & & & & & & & & & \\
10101010 & & & & & & & & & & & & & & & & & & & & & & & \\
01010101 & & & & & & & & & & & & & & & & & & & & & & & \\
 & 11110000 & & & & & & & & & & & & & & & & & & & & & & \\
 & 00001111 & & & & & & & & & & & & & & & & & & & & & & \\
 & 11001100 & & & & & & & & & & & & & & & & & & & & & & \\
 & 00110011 & & & & & & & & & & & & & & & & & & & & & & \\
 & 10101010 & & & & & & & & & & & & & & & & & & & & & & \\
 & 01010101 & & & & & & & & & & & & & & & & & & & & & & \\
 & & 11110000 & & & & & & & & & & & & & & & & & & & & & \\
 & & 00001111 & & & & & & & & & & & & & & & & & & & & & \\
 & & 11001100 & & & & & & & & & & & & & & & & & & & & & \\
 & & 00110011 & & & & & & & & & & & & & & & & & & & & & \\
 & & 10101010 & & & & & & & & & & & & & & & & & & & & & \\
 & & 01010101 & & & & & & & & & & & & & & & & & & & & & \\
 & & & 11110000 & & & & & & & & & & & & & & & & & & & & \\
 & & & 00001111 & & & & & & & & & & & & & & & & & & & & \\
 & & & 11001100 & & & & & & & & & & & & & & & & & & & & \\
 & & & 00110011 & & & & & & & & & & & & & & & & & & & & \\
 & & & 10101010 & & & & & & & & & & & & & & & & & & & & \\
 & & & 01010101 & & & & & & & & & & & & & & & & & & & & \\
 & & & & 1100 & & & & & & & & & & & & & & & & & & & \\
 & & & & 0011 & & & & & & & & & & & & & & & & & & & \\
 & & & & 1010 & & & & & & & & & & & & & & & & & & & \\
 & & & & 0101 & & & & & & & & & & & & & & & & & & & \\
 & & & & & 1100 & & & & & & & & & & & & & & & & & & \\
 & & & & & 0011 & & & & & & & & & & & & & & & & & & \\
 & & & & & 1010 & & & & & & & & & & & & & & & & & & \\
 & & & & & 0101 & & & & & & & & & & & & & & & & & & \\
 & & & & & & 1100 & & & & & & & & & & & & & & & & & \\
 & & & & & & 0011 & & & & & & & & & & & & & & & & & \\
 & & & & & & 1010 & & & & & & & & & & & & & & & & & \\
 & & & & & & 0101 & & & & & & & & & & & & & & & & & \\
 & & & & & & & 10 & & & & & & & & & & & & & & & & \\
 & & & & & & & 01 & & & & & & & & & & & & & & & & \\
\end{bmatrix} \qquad \begin{matrix} \text{0's} \\ \text{elsewhere} \end{matrix} \qquad (2.23)$$

blood cholesterol changes over time is a logistic curve, then the logit function vector $\underset{\sim}{f}$ obtained as $\underset{\sim}{f} = \underset{\sim\sim}{K}\log_e(\underset{\sim}{F})$ becomes of interest, where $\log_e(\underset{\sim}{F})$ is the vector of logarithms of elements in $\underset{\sim}{F}$, and $\underset{\sim}{K}$ is a 19 x 38 block diagonal matrix with nineteen submatrices of the type $\underset{\sim}{K}_1 = [1\ -1]$ forming the main diagonal. A consistent estimate for the covariance matrix of $\underset{\sim}{F}$ is

$$\underset{\sim}{V}_F = \underset{\sim\sim}{K}\underset{\sim\sim}{D}_F^{-1}\underset{\sim}{A}[\underset{\sim}{V}(\underset{\sim}{p}_G)]\underset{\sim}{A}'\underset{\sim}{D}_F^{-1}\underset{\sim}{K}' \qquad (2.24)$$

where $\underset{\sim}{V}(\underset{\sim}{p}_G)$ is a block diagonal matrix with the matrices $\underset{\sim}{V}_i(\underset{\sim}{p}_i) = \{\underset{\sim}{D}_{\underset{\sim}{p}_i} - \underset{\sim}{p}_i\underset{\sim}{p}_i'\}/n_i$ for i = I1, II1, III1, IV1, I2, I3, I4, I5 respectively representing the main diagonal blocks.

In order to compare the groups with respect to trends in their blood cholesterol, we fit the model given by (2.25) to the vector $\underset{\sim}{f}$ by weighted least squares where $\beta_{0,i}$ and $\beta_{1,i}$ represent the intercept and slope parameters for the

$$E\{\underset{\sim}{f}\} = \underset{\sim\sim}{X}\underset{\sim}{\beta} = \begin{bmatrix} 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 1 & 1 & 0 & 0 & 0 & 0 & 0 & 0 \\ 1 & 2 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 2 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 2 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 1 & 1 \\ 0 & 0 & 0 & 0 & 0 & 0 & 1 & 2 \\ 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 1 & 1 & 0 & 0 & 0 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 1 & 2 & 0 & 0 & 0 & 0 & 0 & 0 \\ 1 & 1 & 0 & 0 & 0 & 0 & 0 & 0 \\ 1 & 2 & 0 & 0 & 0 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \end{bmatrix} \begin{bmatrix} \beta_{0,I} \\ \beta_{1,I} \\ \beta_{0,II} \\ \beta_{1,II} \\ \beta_{0,III} \\ \beta_{1,III} \\ \beta_{0,IV} \\ \beta_{1,IV} \end{bmatrix} \qquad (2.25)$$

growth curve fitted to the i-th group $i = I,II,III,IV$. The resulting estimate $\underset{\sim}{b}$ of $\underset{\sim}{\beta}$ and the estimated standard errors are shown in Table 13. The

TABLE 13

ESTIMATED PARAMETERS AND STANDARD ERRORS
FOR CHOLESTEROL MODEL

Parameter	Estimate	s.e.
$\beta_{0,I}$	-1.400	.165
$\beta_{1,I}$	.549	.124
$\beta_{0,II}$	-1.548	.252
$\beta_{1,II}$	1.570	.207
$\beta_{0,III}$	.045	.221
$\beta_{1,III}$	.335	.170
$\beta_{0,IV}$	.043	.229
$\beta_{1,IV}$	1.441	.249

goodness of fit test for the model(2.25) yields $X^2 = 3.61$ with D.F. = 11. Thus, the model is consistent with the data. Results for tests of hypotheses corresponding to pairwise comparisons of the intercepts and slopes in the respective groups have been determined in a manner analogous to (2.7) and are given in Table 14. Hence, it is apparent that treatments I and III have similar time trends as do II and IV. Also treatments I and II have similar initial effects as

TABLE 14

PAIRWISE COMPARISONS OF PARAMETERS
FOR CHOLESTEROL MODEL (X^2_1)

DIETS COMPARED	INTERCEPTS	SLOPES
I-II	0.24	17.81
I-III	27.39	1.03
I-IV	26.06	10.26
II-III	22.58	21.21
II-IV	21.81	0.16
III-IV	0.00	13.45

do III and IV. However, the other comparisons are statistically significant (α=.01). These conclusions may be summarized by fitting the revised model displayed in (2.26)where β_1^* and β_2^* are the two distinct intercept parameters and

$$\underset{\sim}{E\{f\}} = \underset{\sim R}{X}\underset{\sim}{\beta^*} = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 1 & 0 & 1 & 0 \\ 1 & 0 & 2 & 0 \\ 1 & 0 & 0 & 0 \\ 1 & 0 & 0 & 1 \\ 1 & 0 & 0 & 2 \\ 0 & 1 & 0 & 0 \\ 0 & 1 & 1 & 0 \\ 0 & 1 & 2 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 1 & 0 & 1 \\ 0 & 1 & 0 & 2 \\ 1 & 0 & 0 & 0 \\ 1 & 0 & 1 & 0 \\ 1 & 0 & 0 & 0 \\ 1 & 0 & 2 & 0 \\ 1 & 0 & 1 & 0 \\ 1 & 0 & 2 & 0 \\ 1 & 0 & 0 & 0 \end{bmatrix} \begin{bmatrix} \beta_1^* \\ \beta_2^* \\ \beta_3^* \\ \beta_4^* \end{bmatrix} \qquad (2.26)$$

β_3^* and β_4^* are the two distinct slope parameters. On fitting the model (2.26) by weighted least squares, we obtain estimates $\underset{\sim}{b^*}$ and estimated covariance-matrix $\underset{\sim b^*}{V}$ as shown in (2.27). The goodness of fit test for the model (2.26) yields X^2 = 5.69 with D.F. = 15 which implies that this model also is consistent with the data. Finally, for tests of the hypotheses $H_0 : \beta_1^* = \beta_2^*$ and $H_0 : \beta_3^* = \beta_4^*$, X^2 = 89.74 with D.F. = 1 and X^2 = 66.09 with D.F. = 1 respectively. The results are:

$$\begin{bmatrix} b_1^* \\ b_2^* \\ b_3^* \\ b_4^* \end{bmatrix} = \begin{bmatrix} -1.365 \\ -.065 \\ .485 \\ 1.461 \end{bmatrix} \qquad \underset{\sim}{V}_{b^*} = \begin{bmatrix} 1.426 & .628 & -.700 & -.818 \\ & 1.713 & -.614 & -.564 \\ & & .763 & .445 \\ & & & 1.569 \end{bmatrix} \times 10^{-2} \qquad (2.27)$$

3. STATISTICAL THEORY

The methodology which has been applied in the preceding examples is very similar to that given by GSK for the complete data case. In particular, we assume that there are s populations of elements from which independent random samples of fixed sizes $n_1, n_2, \ldots, n_s$ respectively are selected. The responses of the n_i elements from the i-th population are classified into r_i categories, with n_{ij}, where $j=1,2,\ldots,r_i$, denoting the number of elements classified into the j-th response category for the i-th population. This structure is different from that considered in GSK because it allows the definition of response categories as well as their number (i.e., the r_i) to vary from one population to another.

The vector $\underset{\sim}{n}_i$, where $\underset{\sim}{n}_i' = (n_{i1}, n_{i2}, \ldots, n_{ir_i})$, will be assumed to follow the multinomial distribution with parameters n_i and $\underset{\sim}{\pi}_i$, where π_{ij} is the probability of response j in population i. Thus, the relevant product multinomial model is

$$\phi = \prod_{i=1}^{s} \left\{ \frac{n_i!}{\prod_{j=1}^{r_i} n_{ij}!} \prod_{j=1}^{r_i} \pi_{ij}^{n_{ij}} \right\} \qquad (3.1)$$

Other models like those considered by Johnson and Koch [1970] are applicable for more complex sampling situations.

Let $\underset{\sim}{p}_i = (\underset{\sim}{n}_i/n_i)$ and let $\underset{\sim}{p}_G$ be the compound vector defined by $\underset{\sim}{p}_G' = (\underset{\sim}{p}_1', \underset{\sim}{p}_2', \ldots, \underset{\sim}{p}_s')$. A consistent estimate for the covariance matrix of $\underset{\sim}{p}_G$ is given by the block diagonal matrix $\underset{\sim}{V}(\underset{\sim}{p}_G)$ with the matrices $\underset{\sim}{V}_i(\underset{\sim}{p}_i) = \frac{1}{n_i}\{D_{\underset{\sim}{p}_i} - \underset{\sim}{p}_i\underset{\sim}{p}_i'\}$ for i=1,2,...s representing the main diagonal; here $D_{\underset{\sim}{p}_i}$ is a diagonal matrix with elements in the vector $\underset{\sim}{p}_i$ on the main diagonal.

Let $F_1(\underset{\sim}{p}_G), F_2(\underset{\sim}{p}_G), \ldots, F_u(\underset{\sim}{p}_G)$ be a set of u functions of $\underset{\sim}{p}_G$, each with partial derivatives up to order two with respect to the elements of $\underset{\sim}{p}_G$ existing throughout a region containing $\underset{\sim}{\pi}_G = \underset{\sim}{E}\{\underset{\sim}{p}_G\}$ where $\underset{\sim}{\pi}_G' = (\underset{\sim}{\pi}_1', \underset{\sim}{\pi}_2', \ldots, \underset{\sim}{\pi}_s')$. If $\underset{\sim}{F} \equiv \underset{\sim}{F}(\underset{\sim}{p}_G)$ is defined by $\underset{\sim}{F}' \equiv \underset{\sim}{F}'(\underset{\sim}{p}_G) = (F_1(\underset{\sim}{p}_G), F_2(\underset{\sim}{p}_G), \ldots, F_u(\underset{\sim}{p}_G))$, then the covariance matrix of $\underset{\sim}{F}$ can be consistently estimated by $\underset{\sim}{V}_{\underset{\sim}{F}} = \underset{\sim}{H}[\underset{\sim}{V}(\underset{\sim}{p}_G)]\underset{\sim}{H}'$ where $\underset{\sim}{H} = [d\underset{\sim}{F}(\underset{\sim}{x})/d\underset{\sim}{x} \,|_{\underset{\sim}{x}=\underset{\sim}{p}_G}]$. In all applications, the functions comprising $\underset{\sim}{F}$ are chosen so that $\underset{\sim}{V}_{\underset{\sim}{F}}$ is asymptotically non-singular.

The function vector $\underset{\sim}{F}$ is a consistent estimator of $\underset{\sim}{F}(\underset{\sim}{\pi}_G)$. Hence, consideration can then be directed at fitting a linear model $\underset{\sim}{E}\{\underset{\sim}{F}(\underset{\sim}{p}_G)\} \doteq \underset{\sim}{F}(\underset{\sim}{\pi}_G) = \underset{\sim}{X}\underset{\sim}{\beta}$, where $\underset{\sim}{X}$ is a known (u x t) coefficient matrix and $\underset{\sim}{\beta}$ is an unknown (t x 1) parameter vector. Weighted least squares is applied to determine a BAN estimator $\underset{\sim}{b}$ for $\underset{\sim}{\beta}$ as indicated in (2.4). Statistical tests for the fit of the model and for linear hypotheses involving $\underset{\sim}{\beta}$ can be undertaken by applying (2.6) and (2.7).

For the examples considered in Section 2, primary emphasis was placed on cases where $\underset{\sim}{F}(\underset{\sim}{p}_G)$ were linear functions obtained as $\underset{\sim}{F}(\underset{\sim}{p}_G) = \underset{\sim}{A}\underset{\sim}{p}_G$; in this event $\underset{\sim}{V}_{\underset{\sim}{F}} = \underset{\sim}{A}[\underset{\sim}{V}(\underset{\sim}{p}_G)]\underset{\sim}{A}'$. The functions comprising $\underset{\sim}{F}$ in these examples were either estimates of cell probabilities or marginals thereof. Thus, even though the response categories for the respective s populations were not necessarily the same, they were always related in the sense of involving the same underlying parameters. This fact defines the underlying principle by means of which information on both completely classified subjects and incompletely

classified subjects can be synthesized together in the analysis. Finally, it should be indicated that in certain situations like that illustrated in Section 2.3, a two-stage approach is required. At the first stage, the methods which have been described here are used to determine estimates $\hat{\underset{\sim}{\pi}}$ of $\underset{\sim}{\pi}$ which reflect all the available data. Corresponding to $\hat{\underset{\sim}{\pi}}$, an estimated covariance matrix $\hat{\underset{\sim}{V}}$ is also determined. Then, at the second stage, $\hat{\underset{\sim}{\pi}}$ and $\hat{\underset{\sim}{V}}$ play the same role as $\underset{\sim}{p}_G$ and $\underset{\sim}{V}(\underset{\sim}{p}_G)$ do at the first stage. In other words, if $\underset{\sim}{f}(\underset{\sim}{\pi})$ represents a set of functions which are of interest, then corresponding estimates are given by $\underset{\sim}{f}(\hat{\underset{\sim}{\pi}})$ with estimated covariance matrix $\underset{\sim}{V}_{\underset{\sim}{f}} = \tilde{\underset{\sim}{H}}\underset{\sim}{V}\tilde{\underset{\sim}{H}}'$ where $\tilde{\underset{\sim}{H}} = [\underset{\sim}{df}(\underset{\sim}{x})/\underset{\sim}{dx}|_{\underset{\sim}{x}=\hat{\underset{\sim}{\pi}}}]$. Finally, suitable linear models $\underset{\sim}{E}\{\underset{\sim}{f}(\hat{\underset{\sim}{\pi}})\} \doteq \underset{\sim}{f}(\underset{\sim}{\pi}) = \underset{\sim}{X}_{\underset{\sim}{f}}\underset{\sim}{\gamma}$ are then fitted by weighted least squares.

Essentially all the analyses described in this paper can be undertaken by using the same computer program cited in GSK. The only refinements required are suitable modifications of the relevant matrix operations. Also, in some cases, if certain n_{ij} are zero, it may be necessary to replace them by $(1/r_i n_i)$ in order to prevent $\underset{\sim}{V}_{\underset{\sim}{F}}$ from being singular (See GSK and Berkson [1955] for more details in this respect).

In summary, a straightforward generalization of the GSK approach has been formulated for situations involving incompletely classified categorical data. The corresponding methodology was shown to provide the analyst with a powerful tool for dealing with a wide variety of problems. Thus, in some sense, one can say that incomplete categorical data can be more readily handled than incomplete continuous data provided that the sample size is large (e.g., as in the examples reported here). In this context, Kleinbaum [1970] discusses how Wald statistics can be applied to incomplete continuous data arising in

situations like those emphasized here. Alternative methods have been discussed by Afifi and Elashoff [1966, 1967,1969a, 1969b],Hocking and Smith [1968], and Hartley and Hocking [1971]. However, if computational difficulties arise in the application of any of these methods, then it may be more convenient to categorize the continuous data corresponding to such cases and apply the analysis reported here.

REFERENCES

Addelman, S. [1969]. Sequences of two level fractional factorial plans. Technometrics 11, 477-510.

Afifi, A.A. and Elashoff, R.M. [1966]. Missing observations in multivariate statistics I. Review of the literature. J. Amer. Statist. Ass. 61, 595-605.

Afifi, A.A. and Elashoff, R.M. [1967]. Missing observations in multivariate statistics II. Point estimation in simple linear regression. J. Amer. Statist. Ass. 62, 10-29.

Afifi, A.A. and Elashoff, R.M. [1969a]. Missing observations in multivariate statistics III. Large sample analysis of simple linear regression. J. Amer. Statist. Ass. 64, 337-58.

Afifi, A.A. and Elashoff, R.M. [1969b]. Missing observations in multivariate statistics IV. A note on simple linear regression. J. Amer. Statist. Ass. 64, 359-65.

Allen, D.M. and Grizzle, J.E. [1969]. Analysis of growth and dose response curves. Biometrics 25, 357-82.

Berkson, J. [1955]. Maximum likelihood and minimum χ^2 estimates of the logistic function. J. Amer. Statist. Ass. 50, 130-62.

Bhapkar, V.P. [1965]. Categorical data analogs of some multivariate tests. S.N. Roy Memorial Volume, University of North Carolina Press, Chapel Hill, N.C.

Bhapkar, V.P. [1966]. A note on the equivalence of two test criteria for hypotheses in categorical data. J. Amer. Statist. Ass. 61, 228-35.

Bhapkar, V.P. and Koch, G.G. [1968a]. Hypotheses of 'no interaction' in multi-dimensional contingency tables. Technometrics 10, 107-23.

Bhapkar, V.P. and Koch, G.G. [1968b]. On the hypotheses of 'no interaction' in contingency tables. Biometrics 24, 567-94.

Bishop, Y.M.M. [1969]. Full contingency tables, logits, and split contingency tables. Biometrics 25, 383-99.

Bishop, Y.M.M. [1971]. Effects of collapsing multidimensional contingency tables. Biometrics 27, 545-62.

Bishop, Y.M.M. and Fienberg, S.E. [1969]. Incomplete two-dimensional contingency tables. Biometrics 25, 119-28.

Blumenthal, S. [1968]. Multinomial sampling with partially categorized data. J. Amer. Statist. Ass. 63, 542-51.

Box, G.E.P. [1966]. A note on augmented designs. Technometrics 8, 184-88.

Box, G.E.P. and Wilson, K.B. [1951]. On the experimental attainment of optimum conditions. J.R. Statist. Soc. B 13, 1-45.

Cochran, W.G. [1950]. The comparison of percentages in matched samples. Biometrika 37, 256-66.

Daniel, C. [1962]. Sequences of fractional replicates in the 2^{p-q} series. J. Amer. Statist. Ass. 57, 403-29.

Dyke, G.V. and Patterson, H.D. [1952]. Analysis of factorial arrangements when the data are proportions. Biometrics 8, 1-12.

Fienberg, S.E. [1970]. An iterative procedure for estimation in contingency tables. Ann. Math. Statist. 41, 901-17.

Gaylor, D.W. and Merrill, J.A. [1968]. Augmenting existing data in multiple regression. Technometrics 10, 73-82.

Goodman, L.A. [1968]. The analysis of cross-classified data: Independence, quasi-independence, and interactions in contingency tables with or without missing entries. J. Amer. Statist. Ass. 63, 1091-131.

Goodman, L.A. [1970]. The multivariate analysis of qualitative data: interactions among multiple classifications. J. Amer. Statist. Ass. 65, 226-56.

Goodman, L.A. [1971a]. The partitioning of chi-square, the analysis of marginal contingency tables, and the estimation of expected frequencies in multidimensional contingency tables. J. Amer. Statist. Ass. 66, 339-44.

Goodman, L.A. [1971b]. The analysis of multidimensional contingency tables: Stepwise procedures and direct estimation methods for building models for multiple classifications. Technometrics 13, 33-61.

Grizzle, J.E., Starmer, C.F., and Koch, G.G. [1969]. Analysis of categorical data by linear models. Biometrics 25, 489-504.

Hartley, H.O. and Hocking, R.R. [1971]. The analysis of incomplete data. Biometrics 27, 783-823.

Hocking, R.R. and Oxspring, H.H. [1971]. Maximum likelihood estimation with incomplete multinomial data. J. Amer. Statist. Ass. 66, 65-70.

Hocking, R.R. and Smith, W.B. [1968]. Estimation of parameters in the multivariate normal distribution with missing observations. J. Amer. Statist. Ass. 63, 159-73.

Ireland, C.T., Ku, H.H., and Kullback, S. [1969]. Symmetry and marginal homogeneity of an r x r contingency table. J. Amer. Statist. Ass. 64,1323-41.

John, P.W.M. [1966]. Augmenting 2^{n-1} designs. Technometrics 8, 469-80.

Johnson, W.D. and Koch, G.G. [1970]. Analysis of qualitative data. Health Services Research, Winter 1970, 358-69.

Kleinbaum, D.G. [1970]. Estimation and hypothesis testing for generalized multivariate linear models. Unpublished Ph.D. Thesis. University of North Carolina at Chapel Hill. (cf. Mimeo 669).

Koch, G.G. and Reinfurt, D.W. [1970]. The analysis of complex contingency table data from general experimental designs and sample surveys. Proceedings of the Sixteenth Conference on the Design of Experiments in Army Research, Development and Testing, Fort Lee, Va.

Koch, G.G. and Reinfurt, D.W. [1971]. The analysis of categorical data from mixed models. Biometrics 27, 157-74.

Ku, H.H., Varner, R., and Kullback, S. [1968]. Analysis of multidimensional contingency tables. Proceedings of the Fourteenth Conference on the Design of Experiments in Army Research, Development and Testing, Edgewood Arsenal, Md.

Ku,H.H.,Varner, R., and Kullback, S. [1971]. Analysis of multidimensional contingency tables, J. Amer. Statist. Ass. 66, 55-64.

Lewis, J.A. [1968]. A program to fit constants to multiway tables of quantitative and quantal data. Applied Statistics 17, 33-42.

Mantel, N. [1966]. Models for complex contingency tables and polychotomous response curves. Biometrics 22, 83-95.

Mantel, N. [1970]. Incomplete contingency tables. Biometrics 26, 291-304.

Neyman, J. [1949]. Contribution to the theory of the χ^2 test. Pp. 239-73 in Proceedings of the Berkeley Symposium on Mathematical Statistics and Probability, University of California Press, Berkeley and Los Angeles.

Potthoff, R.F. and Roy, S.N. [1964]. A generalized multivariate analysis of variance model useful especially for growth curve problems. Biometrika 51, 122-27.

Reinfurt, D.W. [1970]. The analysis of categorical data with supplemented margins including applications to mixed models. Unpublished Ph.D. Thesis. North Carolina State University, Raleigh. (cf. Mimeo 697).

Roy, S.N., Gnanadesikan, R., Srivastava, J.N. [1971]. Analysis of certain quantitative multiresponse experiments. Pergamon Press, New York.

Roy, S.N. and Kastenbaum, M.A. [1956]. On the hypothesis of no interaction in a multiway contingency table. Ann.Math. Statist. 27, 749-57.

Stuart, A. [1955]. A test for homogeneity of the marginal distributions in a two-way classification. Biometrika 42, 412-16.

Wald, A. [1943]. Tests of statistical hypotheses concerning several parameters when the number of observations is large. Trans. Amer. Math. Soc. 54, 426-82.

Williams, O.D. and Grizzle, J.E. [1970]. Analysis of categorical data with more than one response variable by linear models. University of North Carolina, Institute of Statistics Mimeo Series No. 715.