A Review of Some Statistical Methods for Covariance Analysis of Categorical Data

Gary G. Koch; Ingrid A. Amara; Gordon W. Davis; Dennis B. Gillings

Biometrics, Vol. 38, No. 3, Special Issue: Analysis of Covariance (Sep., 1982), 563-595.

Stable URL:
http://links.jstor.org/sici?sici=0006-341X%28198209%2938%3A3%3C563%3AAROSSM%3E2.0.CO%3B2-Z

Biometrics is currently published by International Biometric Society.

BIOMETRICS 38, 563–595
September 1982

A Review of Some Statistical Methods for Covariance Analysis of Categorical Data

Gary G. Koch, Ingrid A. Amara, Gordon W. Davis and Dennis B. Gillings
Department of Biostatistics, University of North Carolina, Chapel Hill, North Carolina 27514, U.S.A.

SUMMARY

Three general methods for covariance analysis of categorical data are reviewed and applied to an example from a clinical trial in rheumatoid arthritis. The three methods considered are randomization-model nonparametric procedures, maximum likelihood logistic regression, and weighted least squares analysis of correlated marginal functions. A fourth heuristic approach, the unweighted linear model analysis, is an approximate procedure but it is easy to implement. The assumptions and statistical issues for each method are discussed so as to emphasize philosophical differences between their rationales. Attention is given to computational differences, but it is shown that the methods lead to similar results for analogous problems. It is argued that the essential differences between the methods lie in their underlying assumptions and the generality of the conclusions which may be drawn.

1. Introduction

Covariance analysis has become a standard statistical tool and it is used for several different reasons (see Snedecor and Cochran, 1980, Chapter 18). These may be summarized as follows:

(i) To adjust for inherent differences among comparison groups so that bias may be reduced. This is particularly important in observational studies where the equivalence of comparison groups cannot be controlled by the investigator by randomization or matching.

(ii) To generate more powerful statistical tests through variance reduction, which will take place if an appropriate covariable explains some of the variation present. For example, covariance analysis can be employed to advantage in randomized experiments when baseline measurements, taken before the application of treatments, are strongly associated with responses.

(iii) To induce equivalence of comparison groups that are generated by randomization. Equivalence is likely to be approximate as a direct result of randomization and so statistical adjustment can operate to offset minor imbalances for important covariables that are related to the dependent variable of interest.

(iv) To clarify the degree to which treatment effects are explained by other factors. Such explanations would change the interpretation of treatment effects that were otherwise significant. However, the lack of explaining factors would help to substantiate the independent existence of treatment effects.

Key words: Randomization model; Maximum likelihood; Logistic regression; Log-linear model; Weighted least squares; Adjustment; Multivariate normal; Product-multiple hypergeometric; Product-binomial; Product-multinomial; Scores; Rheumatoid arthritis.

(v) To study the degree to which findings are uniform across subpopulations. For example if treatment effects are felt to be specific to certain age groups, an assessment of treatment × age interaction would help to clarify whether treatment effects were generalizable to all ages.

With reference to these reasons, three general methods for covariance analysis of categorical data are contrasted and discussed in terms of the sampling frameworks which motivate them. The general methods are

(i) randomization model nonparametric tests;
(ii) maximum likelihood logistic regression; and
(iii) weighted least squares analysis of correlated marginal functions.

A fourth method, using unweighted linear model analysis, is included for completeness. For each method, the assumptions and computational procedures are discussed for a randomized clinical trial on patients with rheumatoid arthritis. Data for 59 female patients who participated are given in Table 1; patient response status (PRS) is an outcome, or

Table 1
Data listing for rheumatoid arthritis clinical trial

TGRP	Patient number	Age (yrs)	PRS	TGRP	Patient number	Age (yrs)	PRS
1	1	23	5	2	31	31	3
1	2	32	5	2	32	32	5
1	3	37	3	2	33	33	2
1	4	41	2	2	34	37	5
1	5	41	4	2	35	44	5
1	6	48	2	2	36	45	5
1	7	48	5	2	37	46	5
1	8	55	1	2	38	48	4
1	9	55	2	2	39	49	5
1	10	56	2	2	40	51	5
1	11	57	2	2	41	53	5
1	12	57	2	2	42	54	2
1	13	57	2	2	43	54	5
1	14	58	5	2	44	54	5
1	15	59	2	2	45	55	2
1	16	59	1	2	46	57	3
1	17	60	1	2	47	57	4
1	18	61	2	2	48	58	3
1	19	62	2	2	49	59	1
1	20	62	3	2	50	59	3
1	21	66	1	2	51	61	4
1	22	67	2	2	52	63	3
1	23	68	3	2	53	64	5
1	24	68	1	2	54	65	1
1	25	69	3	2	55	66	3
1	26	69	5	2	56	66	4
1	27	70	3	2	57	66	5
2	28	23	4	2	58	68	3
2	29	30	4	2	59	74	2
2	30	30	4				

TGRP: 1, active treatment; 2, placebo treatment.
PRS (Patient response status): 1, Excellent; 2, Good; 3, Moderate; 4; Fair; 5, Poor.

dependent variable, of interest; patient age is an important covariable as arthritis is strongly related to age; and treatment (active vs placebo) is the comparison variable to be evaluated.

The questions of interest for the data in Table 1 are

(i) the extent to which patients on active treatment had more favorable response status than those on placebo;
(ii) the extent to which a treatment difference was maintained after adjustment for age; and
(iii) the extent to which a treatment difference was homogeneous across patient ages and so generalizable to all ages.

These questions can be addressed in ways which correspond to alternative points of view toward the sampling processes generating the data. Accordingly, differences among the three methods of covariance analysis are partly a matter of generalizability of conclusions as justified from the presumed sampling framework, and partly a matter of the underlying reasons for applying a covariance analysis technique.

In §2, the subjects under study are viewed as a finite population which has been randomized into two groups prior to treatment. For this sampling framework, hypergeometric models are used as the basis for unadjusted and covariance-adjusted randomization tests of no difference between treatment effects. The principal advantages of these methods are that they do not require any assumptions about the extent to which subjects are a random sample of some larger population, nor do they involve assumptions about the relationship between patient response status and the covariables. Therefore, no preliminary evaluation of linearity or parallel slopes is required. However, equivalence of the distributions of covariables for the comparison groups needs to be demonstrated, but this equivalence is an expected consequence of randomization. The principal limitation of randomization models is that the results apply only to the finite population of patients initially randomized.

In §3, the patients in each treatment group are presumed to be representative of a larger population of patients of similar ages, thus leading to a sampling framework defined by the product-multinomial distribution for response status across age × treatment subpopulations. Log-linear and logistic models for the relationship of the response distribution to age and treatment are estimated by maximum likelihood methods and then used for covariance-adjusted tests of no difference between treatment effects. The advantage of this approach is that conclusions concerning treatment effects are generalizable across ages. It does not require the covariable distributions to be the same for each treatment group since the covariables correspond here to population characteristics that are fixed and for which subject variation in response is incorporated in the log-linear model. On the other hand, the limitations of the methods of §3 are that subjects may not be representative of any identifiable population and the statistical models may not be compatible with the data according to goodness-of-fit criteria. The use of these methods, therefore, *does* require preliminary evaluation of explicit assumptions such as linearity and parallelism. The methods of §3 are restricted to situations in which the units of study are individual subjects and for which attention is directed at a single response variable. For more general situations, as in studies with repeated-measurements, clusters of subjects and/or multivariate response profiles, it may not be feasible to use realistic models which encompass the overall distribution of the data, but estimation of summary measures for the pattern of variation and their corresponding covariance structure may be straightforward (Koch, Gillings and Stokes, 1980).

The previous considerations motivate the third approach to analysis of covariance, described in §4. Weighted least squares methods are used to fit models which express equivalence of treatment groups for summary measures of the covariables such as mean age; thereby, they provide a framework for adjusted comparisons with respect to response variables. The analysis strategy discussed in §4 is similar to the randomization methods for finite populations in §2 in the sense of requiring equivalence of the two treatment groups for the covariables, but not requiring verification of linear relationships or parallelism. However, it has broader applicability to actual sampling procedures which involve clusters or multivariate response profiles and presumed sampling processes relative to some target population. For the data in Table 1, the subjects in each treatment group can be presumed equivalent to a correspondingly stratified simple random sample (with replacement) from some large population. Since this sampling framework pertains to the joint distribution of age and response for the subjects under study, methods which are based upon it should be viewed cautiously, particularly if some age ranges seem to be over- (or under-) represented. Nevertheless, the advantages of weighted least squares methodology are the broad scope with respect to sampling processes and minimal assumptions concerning the relationship between the response variables and covariables. The limitations are that moderately large samples are required for asymptotic properties to hold, and that conclusions concerning treatment comparisons can only be interpreted in an average sense rather than being generalizable across all specific ages.

Finally, in §5, unweighted least squares covariance analysis is noted to be a convenient computational strategy for approximating the results of the methods in §2, §3 and §4. In this regard, its usefulness with respect to covariance-adjusted tests of significance and the estimation of adjusted means is emphasized. However, caution concerning its application is also recommended.

In summary, covariance analysis of categorical data can be undertaken in several different ways. The principles involved are the same as for continuous data, and the methods analogous, once the sampling processes have been specified.

2. Randomization-Model Methods

Randomization-model methods make no assumptions other than strict randomization of patients to treatment groups and common management and data collection procedures for all patients. The set of 59 patients in Table 1 is viewed as a finite population which has been randomly allocated to two subgroups. Statistical tests of the equivalence of these subgroups can be formulated by applying the theory of simple random sampling (Cochran, 1977, Chapter 2) and arguing that asymptotic chi square approximations are realistic for quadratic forms $Q = \mathbf{y}'\mathbf{V}^{-1}\mathbf{y}$, where $\mathbf{y}$ is a vector of linear sample statistics with mean vector $\mathbf{0}$ and known covariance matrix $\mathbf{V}$ under the null hypothesis of no treatment effects. In §2.1, such test statistics are given for comparisons using four measures of patient response; these are the distribution among the categories 'excellent', 'good', 'moderate', 'fair' and 'poor'; the mean integer score from the scaling (1, 2, 3, 4, 5) for the respective response categories; the mean midrank score from that scaling; and the proportion of subjects with good or excellent response. Covariance-adjusted tests for these four aspects of patient response are given in §2.2 for two types of covariables; these are actual age of patient to illustrate a single covariable, and age group according to four age ranges as an example of covariables with three indicator functions. In §2.3, the results of post-stratification age-adjusted tests are given for comparison purposes.

The principal advantage of the methods in this section is that they yield design-based

inferences; i.e. inferences for which the underlying probability framework follows from randomization as opposed to assumed distributional forms (e.g. binomial, normal, Poisson, etc.) in a spirit similar to that given in Kempthorne (1955); also see Koch and Gillings (1981). Thus, their use is of interest where assumptions might be subject to debate. Also, the covariance methods in §2.2 permit treatment comparisons to be undertaken in a way which adjusts the groups to the same status for the covariables even though their initially-randomized status might have been different. For this reason, differences which are detected by such methods can be attributed more definitively to treatments rather than being partially due to random imbalances for the covariables between the comparison groups. Moreover, this framework for analysis can be interpreted as analogous to that achieved with adaptive stratification designs as discussed in Pocock (1979) and Simon (1979). Since these latter designs involve treatment-assignment probabilities which vary for successive patients in a manner which seeks to maximize marginal balance of the treatment groups for the covariables, their implementation is not straightforward; and so their use may be less practical than simple randomization with covariance adjustment as discussed in §2.2. Otherwise, a capability for more powerful tests concerning treatment-group differences is available with covariance analysis in a randomization framework than without it via a more precise covariance structure.

Finally, although the emphasis throughout this section will be on applications to randomized studies, the methods discussed in it are sometimes applicable to observational studies without randomization. In these situations, randomization is expressed as a hypothesis of interest in its own right and can be used as a basis for analysis. Other aspects of such analysis are beyond the scope of this paper; they are discussed by Koch *et al.* (1980).

2.1 *Direct (Unadjusted) Analysis*

The observed distributions of patient response for the active and placebo treatment groups are summarized in Table 2. The principal question of interest is whether there is any association between patient response status and treatment. More specifically, if there are no differences between treatments, then there should be no association between the assignment to treatment and response status. This implies that the observed values for each patient would be expected to be the same as they would be if assignment had been to the other group; and so the observed values for each treatment group are (by design) a

Table 2
Distributions of patient responses by treatment (with proportions in parentheses)

Treatment	Patient response status					Number of patients
	Excellent, 1	Good, 2	Moderate, 3	Fair, 4	Poor, 5	
Active	5	11	5	1	5	27
	(.18)	(.41)	(.18)	(.04)	(.18)	(1.00)
Placebo	2	4	7	7	12	32
	(.06)	(.12)	(.22)	(.22)	(.38)	(1.00)
Total	7	15	12	8	17	59

simple random sample from the finite population corresponding to their pooled combination as shown in the bottom margin of Table 2. Accordingly, the hypothesis of no association can be expressed as follows:

H_{01}: The active and placebo treatment groups have equivalent distributions of patient response status, compatible with equally likely realizations for the 59!/27!32! possible random partitions of the 59 patients under study. (2.1)

Here, it should be recognized that the statement of H_{01} in (2.1), and other hypotheses and methods discussed in the remainder of this section, apply in the specific form given only to studies like that of Table 1 for which treatment assignment was by simple random sampling; for those using unequally-weighted randomizations such as Efron's (1971) biased-coin design, a different formulation is necessary; see Matts and McHugh (1978) for discussion.

Under the hypothesis H_{01}, the patient responses in Table 2 have the multiple hypergeometric distribution

$$\text{pr}[\{n_{ij}\}] = \prod_{i=1}^{2} n_i! \prod_{j=1}^{5} n_{+j}! \Big/ n! \prod_{i=1}^{2} \prod_{j=1}^{5} n_{ij}!, \tag{2.2}$$

where the $\{n_{ij}\}$ are the numbers of subjects in the active ($i=1$) and placebo ($i=2$) treatment groups with 'excellent' ($j=1$), 'good' ($j=2$), 'moderate' ($j=3$), 'fair' ($j=4$), and 'poor' ($j=5$) response status; the $\{n_i = \sum_{j=1}^{5} n_{ij}\} = (27, 32)$ are the sample sizes for the $s=2$ treatment groups, and the $\{n_{+j} = \sum_{i=1}^{2} n_{ij}\} = (7, 15, 12, 8, 17)$ are the frequencies for the $r=5$ levels of the pooled response distributions of the $n=59$ patients. Both the $\{n_i\}$ and the $\{n_{+j}\}$ marginal distributions are considered fixed, the former by the research design and the latter by the hypothesis H_{01}. From (2.2), it follows that the frequencies $\{n_{ij}\}$ have expected values and covariance structure

$$\begin{aligned} \text{E}(n_{ij} \mid H_{01}) &= n_i n_{+j}/n = m_{ij}, \\ \text{cov}(n_{ij}, n_{i'j'} \mid H_{01}) &= m_{ij}(n\delta_{ii'} - n_{i'})(n\delta_{jj'} - n_{+j'})/n(n-1), \end{aligned} \tag{2.3}$$

where $\delta_{kk'} = 1$ if $k = k'$ and $\delta_{kk'} = 0$ if $k \neq k'$. Since most of the expected frequencies m_{ij} are ≥ 5.0, it is reasonable to regard the $\{n_{ij}\}$ as having a four-dimensional (singular) multivariate normal distribution (since there are six linear restrictions corresponding to their fixed row and column marginal totals). From relevant central limit theory (Puri and Sen 1971), it follows that the quadratic form statistic Q_R which contrasts the $\{n_{ij}\}$ with their expected values $\{m_{ij}\}$ on H_{01}, relative to their covariance structure on H_{01}, has an approximate χ^2 distribution with 4 df. More specifically, this randomization χ^2 statistic can be expressed as

$$\begin{aligned} Q_R &= (\mathbf{n}-\mathbf{m})'\mathbf{A}_R'(\mathbf{A}_R\mathbf{V}\mathbf{A}_R')^{-1}\mathbf{A}_R(\mathbf{n}-\mathbf{m}) = \frac{(n-1)}{n}\sum_{i=1}^{2}\sum_{j=1}^{5}\frac{(n_{ij}-m_{ij})^2}{m_{ij}} \\ &= (n-1)Q_P/n = 11.73, \end{aligned} \tag{2.4}$$

where $\mathbf{n} = (5, 11, 5, 1, 5, 2, 4, 7, 7, 12)'$ is the compound vector of observed frequencies; $\mathbf{m}$ and $\mathbf{V}$ are its corresponding expected frequency vector and covariance matrix from (2.3), and $\mathbf{A}_R = [\mathbf{I}_4, \mathbf{0}_{4,6}]$ is a 4×10 matrix which forms observed minus expected frequency differences for the excellent, good, moderate, and fair response levels for the active treatment, via the 4×4 identity matrix $\mathbf{I}_4$ and the 4×6 null matrix $\mathbf{0}_{4,6}$; also, Q_P is the well-known Pearson chi square statistic for testing H_{01}. The identity $Q_P = nQ_R/(n-1)$ can

be verified by matrix arguments like those given in Koch and Bhapkar (1981). Since Q_R in (2.4) is statistically significant with $P<.05$, the hypothesis H_{01} is contradicted. Examination of Table 2 indicates that active group patients are more likely to achieve good or excellent response status while those in the placebo group are more likely to have fair or poor response status.

Another method which can be used to test H_{01} is the one-way analysis of variance randomization criterion

$$Q_{RS} = (\mathbf{n}-\mathbf{m})'\mathbf{A}'_{RS}(\mathbf{A}_{RS}\mathbf{V}\mathbf{A}'_{RS})^{-1}\mathbf{A}_{RS}(\mathbf{n}-\mathbf{m}) = (n-1)\sum_{i=1}^{2} n_i(\bar{a}_i-\bar{a})^2/nv_a$$
$$= (n-1)S_a^2/S_t^2, \tag{2.5}$$

where $\mathbf{a}=(a_1, a_2, a_3, a_4, a_5)'$ is a set of ordinal scores reflecting response status; the $\bar{a}_i=\sum_{j=1}^{5} a_j n_{ij}/n_i$ are corresponding sample means for the two treatment groups; $\bar{a}=(\sum_{j=1}^{5} a_j n_{+j}/n)$ and $v_a=\sum_{j=1}^{5}(a_j-\bar{a})^2(n_{+j}/n)$ are the finite population mean and variance for all subjects under study; and $\mathbf{A}_{RS}=[\mathbf{a}', \mathbf{0}'_{1,5}]$ is a 1×10 matrix which forms a linear combination of the observed minus expected frequency differences for the respective response levels for the active treatment with coefficients a_1, a_2, a_3, a_4 and a_5; $S_a^2=\sum_{i=1}^{2} n_i(\bar{a}_i-\bar{a})^2$ is the among-groups sum of squares; and $S_t^2=nv_a$ is the total sum of squares. As before, the identities in (2.5) can be verified by the use of matrix arguments like those given in Koch and Bhapkar (1982).

Each of the two forms of Q_{RS} given in (2.5) are of interest for different reasons. The matrix expression provides the basis for Q_{RS} having an approximate χ^2 distribution with 1 df as a consequence of its being a quadratic form statistic which contrasts a single function with zero expected value under H_{01} relative to its H_{01} variance; it is also readily implemented in general-purpose computer programs such as PARCAT which is documented by Landis *et al.* (1979). Alternatively, $(n-1)S_a^2/S_t^2$ is convenient for hand calculations and has intuitive appeal.

For the data summarized in Table 2, Q_{RS} has been obtained for three sets of scores $\mathbf{a}$. These were integer scores $\mathbf{a}=(1, 2, 3, 4, 5)'$ for which

$$Q_{RS} = 58[\{27(2.63-3.22)^2+32(3.72-3.22)^2\}/116.14] = 8.68, \tag{2.6}$$

midrank scores $\mathbf{a}=(4, 15, 28.5, 38.5, 51)'$ for which $Q_{RS}=8.73$, and a binary indicator $\mathbf{a}=(1, 1, 0, 0, 0)'$ for good or excellent response for which $Q_{RS}=10.10$. Since the Q_{RS} test statistics are all significant with P-values $<.01$ for χ^2 on 1 df, they provide stronger contradiction of H_{01} than the overall statistic Q_R as χ^2 on 4 df. Thus, they illustrate that Q_{RS} can be more powerful than Q_R through targeting smaller degrees of freedom at location-shift alternatives expressed in terms of expected differences among the mean scores $\{\bar{a}_i\}$. Another advantage of Q_{RS} over Q_R is that its sample-size requirements for χ^2 approximations are less stringent since they refer to the mean scores $\{\bar{a}_i\}$ [which are linear combinations of the $\{n_{ij}\}$] rather than the $\{n_{ij}\}$. Thus, such methods are similar in spirit to those suggested by Cochran (1954) for strengthening contingency table χ^2 tests; this comment is also applicable to the methods described later in this paper. Finally, integer scores are of interest for Q_{RS} if the response categories can be viewed as equally spaced, but also can be useful in general situations for reasons given in Koch *et al.* (1977) and Koch *et al.* (1980); midrank scores are of interest for producing the contingency table version of the Wilcoxon rank sum statistic with no scaling of the response categories other than that implied by their observed relative ordering in the study population; and the binary indicator is of interest as a directly interpretable measure of favorable response.

2.2 *Covariance-Adjusted Analysis*

For many investigations, data are available for one or more covariables which provide background information on sources of subject heterogeneity. Since covariables like age are typically known prior to the administration of treatment, their distributions for the two randomized groups should be equivalent; i.e. the randomization hypothesis H_{02} applies to age in a form analogous to (2.1). Thus, patient response status and age may be viewed as having opposite roles because the detection of significant differences is of interest for the former, while the verification of the nonexistence of such differences is of interest for the latter. These roles can be unified by considering the joint randomization hypothesis

H_{03}: The active and placebo treatment groups have equivalent bivariate distributions of age and response status, compatible with equally likely realizations for the 59!/27!32! possible random partitions of the 59 patients under study. (2.7)

The finite population simple random sampling framework specified by H_{03} implies that the expected values and covariance structure for the mean vectors $\bar{\mathbf{y}}_i$ and $\bar{\mathbf{x}}_i$ of response status and age for the two treatment groups $i = 1, 2$ are

$$\mathrm{E}\left(\begin{bmatrix}\bar{\mathbf{y}}_i\\ \bar{\mathbf{x}}_i\end{bmatrix} \mid H_{03}\right) = \mathrm{E}\left(\frac{1}{n_i}\sum_{l=1}^{n_i}\begin{bmatrix}\mathbf{y}_{il}\\ \mathbf{x}_{il}\end{bmatrix} \mid H_{03}\right) = \begin{bmatrix}\bar{\mathbf{y}}\\ \bar{\mathbf{x}}\end{bmatrix} = \frac{1}{n}\sum_{i=1}^{2}\sum_{l=1}^{n_i}\begin{bmatrix}\mathbf{y}_{il}\\ \mathbf{x}_{il}\end{bmatrix}, \tag{2.8}$$

$$\begin{aligned}\mathrm{cov}\left(\begin{bmatrix}\bar{\mathbf{y}}_i\\ \bar{\mathbf{x}}_i\end{bmatrix}, \begin{bmatrix}\bar{\mathbf{y}}_{i'}\\ \bar{\mathbf{x}}_{i'}\end{bmatrix} \mid H_{03}\right) &= \frac{(n\delta_{ii'} - n_i)}{n_i(n-1)}\begin{bmatrix}\mathbf{V}_y & \mathbf{V}'_{xy}\\ \mathbf{V}_{xy} & \mathbf{V}_x\end{bmatrix}\\ &= \frac{(n\delta_{ii'} - n_i)}{n_i(n-1)n}\sum_{i=1}^{2}\sum_{l=1}^{n_i}\left(\begin{bmatrix}(\mathbf{y}_{il} - \bar{\mathbf{y}})\\ (\mathbf{x}_{il} - \bar{\mathbf{x}})\end{bmatrix}[(\mathbf{y}_{il} - \bar{\mathbf{y}})', (\mathbf{x}_{il} - \bar{\mathbf{x}})']\right),\end{aligned} \tag{2.9}$$

where $\mathbf{y}_{il}$ and $\mathbf{x}_{il}$ denote the randomly-assigned vectors expressing aspects of patient response and age for the lth subject in the ith treatment group, while $n\bar{\mathbf{y}}$ and $n\bar{\mathbf{x}}$ are their fixed finite population totals [see Cochran (1977, Chapter 2) for further explanation concerning random sample values vs fixed population values]. One definition of interest for $\mathbf{y}_{il}$ is the vector $\mathbf{y}_{il} = (y_{i1l}, y_{i2l}, y_{i3l}, y_{i4l})'$ of indicator variables y_{ijl}, which equal 1 if the lth subject in the ith group has the jth response status and equal 0 otherwise, for $j = 1, 2, 3, 4$; note that $y_{i5l} = 1 - \sum_{j=1}^{4} y_{ijl}$. Alternatively, $\mathbf{y}_{il}$ can be a univariate score $a_{il} = \sum_{j=1}^{5} a_j y_{ijl}$ for the lth subject in the ith group for ordinal data situations. Similarly, $\mathbf{x}_{il}$ can either be a vector $\mathbf{x}_{il} = (x_{i1l}, x_{i2l}, x_{i3l})$ of three indicator variables for four age groups or it can be the univariate value x_{i*l} for the subject's actual age.

The sample sizes $n_i = 27, 32$ for this example are considered sufficiently large for the mean vectors $\bar{\mathbf{y}}_i$ and $\bar{\mathbf{x}}_i$ described here to have approximately a joint multivariate normal distribution. Thus, the hypothesis H_{03} can be tested via the multivariate randomization criterion

$$Q_{MR}(\mathbf{y}, \mathbf{x}) = \{(n-1)/n\}\sum_{i=1}^{2} n_i[(\bar{\mathbf{y}}_i - \bar{\mathbf{y}})', (\mathbf{x}_i - \bar{\mathbf{x}})']\begin{bmatrix}\mathbf{V}_y & \mathbf{V}'_{xy}\\ \mathbf{V}_{xy} & \mathbf{V}_x\end{bmatrix}^{-1}\begin{bmatrix}\bar{\mathbf{y}}_i - \bar{\mathbf{y}}\\ \bar{\mathbf{x}}_i - \bar{\mathbf{x}}\end{bmatrix}, \tag{2.10}$$

which has an approximate χ^2 distribution with $d + t$ df, where d is the dimension of the $\mathbf{y}_{il}$ and t is the dimension of the $\mathbf{x}_{il}$; more generally, for s groups, $\mathrm{df} = (s-1)(d+t)$. Also, if the $\mathbf{y}_{il}$ and $\mathbf{x}_{il}$ are vectors of ranks relative to the combined groups (with midranks used for ties), then $Q_{MR}(\mathbf{y}, \mathbf{x})$ is the multivariate Kruskal–Wallis statistic discussed by Puri and Sen (1971).

The statistic $Q_{MR}(\mathbf{y}, \mathbf{x})$ can be partitioned into two independent components under H_{03} as

$$\begin{aligned} Q_{MR}(\mathbf{y}, \mathbf{x}) &= Q_{MR}(\mathbf{x}) + \{Q_{MR}(\mathbf{y}, \mathbf{x}) - Q_{MR}(\mathbf{x})\} \\ &= Q_{MR}(\mathbf{x}) + Q_{MR}(\mathbf{y} \mid \mathbf{x}). \end{aligned} \tag{2.11}$$

The statistic $Q_{MR}(\mathbf{x})$, which has the same form as (2.10) with the $\mathbf{y}$ terms excluded, is the multivariate criterion for testing the randomization hypothesis H_{02} for the covariables. The multivariate statistic $Q_{MR}(\mathbf{y} \mid \mathbf{x})$ can be interpreted as a covariance-adjusted test statistic because it can be shown [by matrix arguments like those in Koch and Bhapkar (1982)] to have the form

$$Q_{MR}(\mathbf{y} \mid \mathbf{x}) = \{(n-1)/n\}\left(\sum_{i=1}^{2} n_i \bar{\mathbf{g}}_i' \mathbf{V}_{\mathbf{g}}^{-1} \bar{\mathbf{g}}_i\right) = Q_{MR}(\mathbf{g}) \equiv Q_{RC}, \tag{2.12}$$

with $\bar{\mathbf{g}}_i = (\bar{\mathbf{y}}_i - \bar{\mathbf{y}}) - \mathbf{V}_{xy}' \mathbf{V}_x^{-1}(\bar{\mathbf{x}}_i - \bar{\mathbf{x}})$ and $\mathbf{V}_g = \mathbf{V}_y - \mathbf{V}_{xy}' \mathbf{V}_x^{-1} \mathbf{V}_{xy}$; also it is assumed that the joint data matrix for response variables and covariables is nonredundant so that $\mathbf{V}_g^{-1}$ exists. Since the criterion $Q_{MR}(\mathbf{y} \mid \mathbf{x})$ is a multivariate randomization statistic like (2.10) with respect to the residuals $\mathbf{g}_{il} = (\mathbf{y}_{il} - \bar{\mathbf{y}}) - \mathbf{V}_{xy}' \mathbf{V}_x^{-1}(\bar{\mathbf{x}}_{il} - \bar{\mathbf{x}})$ of a combined sample set of multiple regressions of the $\mathbf{y}_{il}$ on the $\mathbf{x}_{il}$, it is directly analogous to the rank analysis of covariance methods discussed by Quade (1967). Also, see Quade (1982) concerning the use of matching principles to construct the $\mathbf{g}_{il}$ in a spirit similar to the stratification adjustments considered in §2.3.

Another way in which $Q_{RC} \equiv Q(\mathbf{y} \mid \mathbf{x})$ can be interpreted as a covariance-adjusted statistic is through recognizing $Q_{MR}(\mathbf{x})$ as a goodness-of-fit test statistic for the weighted least squares regression fit of a joint linear model to the $\bar{\mathbf{y}}_i - \bar{\mathbf{y}}$ and $\bar{\mathbf{x}}_i - \bar{\mathbf{x}}$. In this case, the parameters for the $\bar{\mathbf{x}}_i - \bar{\mathbf{x}}$ are all restricted to $\mathbf{0}$ (given H_{02}) so that those for the $\bar{\mathbf{y}}_i - \bar{\mathbf{y}}$ are the resulting adjusted values predicted from the association of the $\mathbf{y}$ terms with the $\mathbf{x}$ terms if the means $\bar{\mathbf{x}}_i$ for the respective samples had actually been the same. In other words, Q_{RC} permits the association between groups and response variables to be assessed in a setting for which the groups have the same status for the covariables even though their initially-randomized status might have been different.

A further point which can be noted is that the term 'analysis of covariance' here means that the covariation of the $\mathbf{y}$ with the $\mathbf{x}$ is used to formulate a test statistic for the $\mathbf{y}$ which reflects the known information that the hypothesis H_{02} applies to the $\mathbf{x}$. As such, it involves a regression adjustment in a spirit similar to the regression estimate of the mean in finite population simple random sampling (see Cochran, 1977, Chapter 7). However, a *prediction model* for the relationship of the y variables with the x variables is not explicitly required. Accordingly, tests of parallelism (or equality of regression coefficients) for the respective groups are not necessary for randomization-model-based covariance analysis because they pertain to the structure of a prediction model and the implications of covariables to the generalizability of conclusions as discussed in §3. However, verification of H_{02} via the criterion $Q_{MR}(\mathbf{x})$ is important. In this regard, its acceptance enhances the rationale for the use of the covariance-adjusted criterion Q_{RC} to test H_{04}; while its contradiction would tend to suggest a real imbalance in the allocation of subjects to groups, which might imply that the randomization model for (2.7)–(2.9) was not a reasonable framework for the analysis of the response variables. In this type of situation, post-stratification might be used to reduce the imbalance so that stratified randomization models as discussed in §2.3 could be used.

A specific application of the randomization covariance statistic Q_{RC} involves integer scores for patient response and actual age in years as the covariable. For this case, the

multivariate randomization criterion (2.10) is

$$Q_{MR}(\text{integer response, age}) = (58/59)\left(27[-0.59, 2.2]\begin{bmatrix} 1.97 & -6.60 \\ -6.60 & 160.8 \end{bmatrix}^{-1}\begin{bmatrix} -0.59 \\ 2.2 \end{bmatrix} + 32[0.50, -1.9]\begin{bmatrix} 1.97 & -6.60 \\ -6.60 & 160.8 \end{bmatrix}^{-1}\begin{bmatrix} 0.50 \\ -1.9 \end{bmatrix}\right) = 8.70; \tag{2.13}$$

and under H_{03}, it has an approximate χ^2 distribution with 2 df. The randomization test statistic for H_{02} with respect to age is

$$Q_R(\text{age}) = (58/59)\{27(2.2)^2/(160.8) + 32(1.9)^2/160.8\} = 1.52, \tag{2.14}$$

and has an approximate χ^2 distribution with 1 df. Since Q_R in (2.14) is nonsignificant with $P \geqslant .10$, it is reasonable to compute the randomization covariance statistic

$$Q_{RC} = Q_{RC}(\text{integer response} \mid \text{age}) = 8.70 - 1.52 = 7.18, \tag{2.15}$$

which under H_{04} has an approximate χ^2 distribution with 1 df. This result is significant with $P < .01$ and thereby contradicts the hypothesis H_{04} in the sense of indicating that the adjusted mean score $\bar{y} + g_1 = \{2.63 + (6.60/160.8)(2.2)\} = 2.72$ for active treatment is more favorable than its counterpart $\bar{y} + g_2 = \{3.72 + (6.60/160.8)(-1.9)\} = 3.64$ for placebo treatment. However, it should be noted that the difference between these adjusted means is smaller than the difference between the unadjusted means of 2.63 and 3.72, and this reduction provides an explanation as to why the covariance statistic Q_{RC} in (2.15) is somewhat smaller than its unadjusted counterpart Q_{RS} in (2.6). This has occurred because patient response status has a negative association with age and there tended to be more younger patients assigned to placebo, and so the placebo group was randomly at some disadvantage with respect to age. The covariance analysis adjusts for this aspect of randomization by having the statistic (2.12) for comparing the groups obtained from a framework in which the mean ages for the two groups are equalized via its implied linear model. For this reason, the tendency for the covariance statistic Q_{RC} to be larger or smaller than its unadjusted counterpart Q_{RS} in any specific study is partly a random event which depends upon whether relatively more or fewer patients with 'less favorable covariable status' are assigned to the more effective treatment. The other important consideration is whether the adjusted covariance matrix $\mathbf{V}_g = \{1.97 - (6.60)^2/160.8\} = 1.70$ in (2.12) is a more sensitive framework for comparison than $\mathbf{V}_y = 1.97$. In the example here, the randomization covariance statistic was found to be smaller than the corresponding direct test statistic. On the other hand, for a crossover design example in Koch and Stokes (1981), it was found to yield a significant result for the Period 1 comparison while the direct test did not.

Results of other types of randomization covariance analyses of the data in Table 1 and also that from (2.15) are given in Table 3. These pertain to selected combinations of four definitions of patient response status (integer scores, midrank scores, the binary indicator for good or excellent, and multiple indicators for excellent, good, moderate and fair) and three definitions of the age covariable (actual age, ranks of age, and age indicators for $\leqslant$44 years, 45–54 years, and 55–64 years). For these analyses, the test statistics for H_{02} of no association for the covariables were all nonsignificant ($P \geqslant .10$); and the covariance adjusted test statistics for H_{04} were all significant ($P < .05$) except for the test for the multiple response indicators adjusted for the age indicators which approached significance with $P = .076$.

Table 3
Randomization chi square tests comparing treatments

Definitions for treatment group comparisons	Direct tests		Covariance-adjusted tests: Actual age		Covariance-adjusted tests: Age groups†		Stratification by age adjusted tests†	
	Q	df	Q	df	Q	df	Q	df
Patient response								
Integer scores, I	8.68 $P = .003$**	1	7.18 $P = .007$**	1	4.15 $P = .042$*	1	4.33 $P = .038$*	1
Midrank scores, R	8.73 $P = .003$**	1	6.76 $P = .009$**	1	4.22 $P = .040$*	1	5.53 $P = .019$*	1
Good or excellent indicator, GE	10.10 $P = .001$**	1	8.79 $P = .003$**	1	6.82 $P = .009$**	1	7.02 $P = .008$**	1
Excellent, good, moderate, fair multiple indicators, M	11.73 $P = .019$*	4	10.21 $P = .037$*	4	8.47 $P = .076$	4	8.73 $P = .068$	4
Age covariables								
Age, A	1.52 $P = .218$	1	Age in years, A, is the actual age covariable for the analyses of I, GE, and M; and midrank age, AR is the covariable for R. The age indicators, AI, are the age-group covariables.				For age stratification, ranks of patient response are within age.	
Midrank age, AR	2.14 $P = .144$	1						
Age indicators for ≤44, 45–54, 55–64, AI	5.47 $P = .140$	3						

* Significant with $.01 < P \leq .05$.
** Significant with $P \leq .01$.
† The age groups for which adjustment is undertaken are (<44, 45–54, 55–64, ≥65).

The principal reason for using the covariance statistic Q_{RC} in a finite-population setting is that it provides a refined comparison of the two treatment groups via adjustments which induce equivalence of randomized differences between them with respect to the covariables. Any subsequent differences among treatment groups are not partially confounded with randomized differences in the distributions of the covariables for the groups. Such test statistics may also yield stronger results than their direct-test counterparts through a more precise covariance structure; but the extent to which they actually do so depends upon the tendency for subjects with less favorable covariable status to be randomly assigned to the more effective treatment.

2.3 *Stratification-Adjusted analysis*

Another way in which information on sources of subject heterogeneity in a study can be used to strengthen the statistical basis for treatment comparisons is through stratification. For the example here, age is a *post-stratification* variable employing the intervals ≤44, 45–54, 55–64 and ≥65 years. The observed distributions of patient response status

Table 4
Patient responses by age and treatment

Age	Treatment	Patient response status					Number of patients
		Excellent, 1	Good, 2	Moderate, 3	Fair, 4	Poor, 5	
≤44	Active	0	1	1	1	2	5
	Placebo	0	1	1	3	3	8
45–54	Active	0	1	0	0	1	2
	Placebo	0	1	0	1	7	9
55–64	Active	3	8	1	0	1	13
	Placebo	1	1	4	2	1	9
≥65	Active	2	1	3	0	1	7
	Placebo	1	1	2	1	1	6

for the active and placebo treatment subgroups of these domains are summarized in Table 4. Alternatively, the age intervals might have been design strata within which subjects were independently randomized to treatments. In that case, restricted randomization would have led to controlled (e.g. equal) numbers of subjects per treatment within each stratum rather than the unrestricted pattern shown in Table 4. However, as noted by Simon (1979), restricted randomization is often not used because it complicates study management.

Regardless of whether stratification is undertaken prior to randomization (by design) or afterwards as an analytic strategy, the hypothesis analogous to H_{01} for no association between patient response status and treatment which takes it into account can be expressed as follows:

H_{05}: For each of the respective age groups ≤44, 45–54, 55–64 and ≥65, the active and placebo treatment groups are equivalent in the sense of H_{01} in (2.1). (2.16)

Under the hypothesis H_{05}, the summary data in Table 4 have the product-multiple hypergeometric distribution

$$\mathrm{pr}[\{n_{hij}\}] = \left\{ \prod_{h=1}^{4} \left(\prod_{i=1}^{2} n_{hi}! \prod_{j=1}^{5} n_{h+j}! \Big/ n_{h+}! \prod_{i=1}^{2} \prod_{j=1}^{5} n_{hij}! \right) \right\}, \tag{2.17}$$

where the $\{n_{hij}\}$ are the frequencies of the jth response status for subjects in the ith subgroup of the ≤44 ($h = 1$), 45–54 ($h = 2$), 55–64 ($h = 3$) and ≥65 ($h = 4$) age domains with $i = 1, 2 = s$ and $j = 1, 2, 3, 4, 5 = r$ having the same definitions as for (2.2) in §2.1; the $\{n_{hi} = \sum_{j=1}^{5} n_{hij}\}$ are the sample sizes for the age × treatment subgroups, the $\{n_{h+j} = \sum_{i=1}^{2} n_{hij}\}$ are the frequencies for the pooled-across-treatments response distributions of the $\{n_{h+} = \sum_{i=1}^{2} \sum_{j=1}^{5} n_{hij}\} = (13, 11, 22, 13\}$ subjects for the respective age domains. The expected values and covariance structure for the $\{n_{hij}\}$ have the same form as given in (2.3) within the hth stratum except that all terms have h as the first subscript, and the covariances for quantities from different strata are all 0.

One method for testing H_{05} involves average partial association statistics which are

directed at the across-strata sums $\sum_{h=1}^{q}(n_{hij}-m_{hij})$. These test criteria, which are extensions of those of Mantel and Haenszel (1959) and Mantel (1963), have the general form

$$Q_{AR}=\left\{\sum_{h=1}^{4}(\mathbf{n}_h-\mathbf{m}_h)'\mathbf{A}_h'\right\}\left\{\sum_{h=1}^{4}\mathbf{A}_h\mathbf{V}_h\mathbf{A}_h'\right\}^{-1}\left\{\sum_{h=1}^{4}\mathbf{A}_h(\mathbf{n}_h-\mathbf{m}_h)\right\}, \tag{2.18}$$

where $\mathbf{n}_h, \mathbf{m}_h$ and $\mathbf{V}_h$ are the observed frequency vector, expected frequency vector and covariance matrix for the hth stratum; and the $\mathbf{A}_h$ are $u\times rs$ matrices which specify $u\leq(r-1)(s-1)=4$ linear functions of the rs response $\times$ treatment categories for the strata $h=1,2,3,4$ in a manner analogous to the $\mathbf{A}$-matrices in (2.4) or (2.5); otherwise, the $\mathbf{A}_h$ are allowed to be different across strata as long as they yield conceptually similar functions such as mean ranks, and they are assumed to be nonredundant so that $\{\sum_{h=1}^{4}\mathbf{A}_h\mathbf{V}_h\mathbf{A}_h'\}$ is nonsingular. The average partial association statistic Q_{AR} in (2.18) has an approximate χ^2 distribution with u df under H_{05} when the across-strata sample sizes $\{n_{+i}=\sum_{h=1}^{4}n_{hi}\}=(27,32)$ are large. It is also a reasonably powerful method for alternatives where the pattern of treatment differences is similar within the strata. However, average partial association statistics should be viewed cautiously when such patterns are conflicting (i.e. in opposite directions), and their use in such situations may be less effective than the methods discussed in §3 and §4. Other discussion concerning average partial association statistics is given in Landis, Heyman and Koch (1978) and Koch *et al.* (1980); extensions which encompass covariance adjustment are described in Amara and Koch (1980) with respect to a general computing procedure.

Results from the application of the average partial association criterion Q_{AR} in (2.18) to the summary data in Table 4 are given in the last column of Table 3 for four definitions of the $\mathbf{A}_h$ matrices. These corresponded to integer scores, within-strata midrank scores, the binary indicator for good or excellent response, and multiple indicators for excellent, good, moderate and fair. Thus, it can be seen that the statistics from the stratification-adjusted analyses are similar to those from the covariance analyses with respect to the age-group indicators with the ones for integer scores, midrank scores, and the good or excellent indicator being significant with $P<.050$ and that for the multiple indicators approaching significance with $P=.068$. Also, these results are not as strong as those for direct tests because they are based upon within-stratum differences for the two treatments and thereby are adjusted for random differences in the age distributions of their respective patients.

Some additional insight concerning stratification adjustment can be gained by noting that for the comparison of two treatments, the statistic Q_{AR} in (2.18) from a single set of scores may be written as

$$Q_{AR}=\left\{\sum_{h=1}^{4}\sum_{j=1}^{5}a_{hj}(n_{hij}-m_{hij})\right\}^2\Big/\left\{\sum_{h=1}^{4}\sum_{j=1}^{5}(a_{hj}-\bar{a}_h)^2n_{h+j}\right\}$$
$$\times\{n_{h1}n_{h2}/n_{h+}(n_{h+}-1)\}=G^2/v_G; \tag{2.19}$$

the $\{\bar{a}_h=\sum_{j=1}^{5}a_{hj}n_{h+j}/n_{h+}\}$ are the finite population means for the respective strata $h=1,2,3,4$ with respect to their scores $\{a_{hj}\}$; G is the function at which the stratification-adjusted test is directed, and v_G is its corresponding variance. In (2.19), G can be interpreted as a linear combination of observed minus expected frequency differences across response levels and age strata for the active treatment. If the $\{a_{hj}\}$ are within-stratum rank scores, then $G=\sum_{h=1}^{q}\{T_{h1}-\frac{1}{2}n_{h1}(n_{h+}+1)\}$ where T_{h1} is the Wilcoxon rank-sum statistic with respect to the first treatment for the hth stratum. In this case, Q_{AR} is the contingency table counterpart of the extension by Benard and van Elteren (1953) of the

stratified rank statistic of Friedman (1937) to the case where the numbers of subjects per treatment in each stratum could be different from one. Similarly, if the $\{a_{hj}\}$ are within-stratum ranks divided by their stratum sizes n_h to form ridits (or divided by the n_h+1), then Q_{AR} is the contingency-table counterpart of the stratified-sample rank statistic discussed by Mack and Skillings (1980), or that of van Elteren (1960).

The extent to which Q_{AR} is larger or smaller than its unadjusted counterpart Q_{RS} in (2.5) depends on two considerations: (i) whether the stratification-based expected frequencies m_{hij} are more or less distant from the n_{hij} as they relate to G^2 than the unadjusted expected frequencies m_{ij} in (2.3) are to the n_{ij} as they relate to S_a^2 in (2.5); (ii) the extent to which the denominator variance v_G for G from stratification is less than its counterpart in (2.5). For the example here, the placebo group had relatively more young patients for whom a less favorable response was more likely, and so stratification adjustment led to expected frequencies which were closer to the corresponding observed frequencies. Thus, the impact of stratification was similar to that of covariance adjustment in §2.2 in the sense that the framework by which each accounted for subject heterogeneity with respect to age involved summary measures which indicated smaller treatment differences than the unadjusted comparison. The principal advantage of such analysis is that a more explicit framework is provided for the attribution of significant differences between the two randomized groups to treatment effects in the spirit of contrasting 'like with like'. Also, it provides the potential for yielding more strongly significant results than unadjusted direct tests via a more precise covariance structure. For further discussion of stratification, including its role as a design strategy and its implications to analysis, see Simon (1979).

3. Log-Linear Model Methods

In this section, the patients in each treatment group are presumed to have responses which are representative of patients of similar age from some large conceptual population of potential patients. More formally, the data are regarded as equivalent to a stratified simple random sample (with replacement) from an age × treatment cross-classification of this population. Since these considerations imply the product-multinomial distribution for response status across age × treatment subpopulations, maximum likelihood methods can be used to fit log-linear models for purposes of covariance analysis. In §3.1, logistic models are investigated for the binary indicator of good or excellent response vs moderate, fair or poor response. These encompass three types of covariables which are indicator functions for four mutually exclusive age ranges, a set of consecutive integers as 'designated midpoint values' for the age ranges, and actual age. Similarly oriented log-linear models for the entire set of five response categories are considered in §3.2, together with a model which takes into account their ordinality.

The methods discussed in this section do not require randomized assignment of treatments, and so they are also applicable to observational studies; see Anderson *et al.* (1980), Cochran (1972) and Snedecor and Cochran (1980, Chapter 18). The stratified simple random sampling assumptions should be viewed carefully because they need to encompass any sources of bias that might contaminate treatment comparisons. Thus, an objective for covariance analysis here is adjustment for sources of bias via models that express the relationship of the response outcome to the corresponding covariables and treatment (each of which might preferably be called 'explanatory' variables). Moreover, these models have the advantage of expressing the extent to which conclusions concerning treatment comparisons can be generalized to all possible values of the covariables. For

this reason, they have broader scope than the methods in §2, but the rationale for their use needs to be supported by statistical and substantive justification of the underlying assumptions about the sampling process and model structure. The basic strategy is that covariables corresponding to sources of bias must always be taken into account regardless of any potential gain in power; but once this is done, others could be added for their implications to variance reduction and greater power, particularly if there is *a priori* reason for doing this.

3.1 *Logistic Model for Binary Outcomes*

The observed frequencies f_{hi1} of good or excellent response and f_{hi2} of moderate, fair or poor response, for the age × treatment cross-classification of patients into $s = 4 \times 2 = 8$ subgroups, are shown in Table 5. Since the data are presumed here to be conceptually representative of some large population in a stratified simple random sampling sense, the frequencies $\{\mathbf{f}_{hi} = (f_{hi1}, f_{hi2})'\}$ have the product-binomial distribution

$$\phi(\mathbf{f} \mid \boldsymbol{\theta}) = \prod_{h=1}^{4} \prod_{i=1}^{2} n_{hi}!\, \theta_{hi}^{f_{hi1}} (1-\theta_{hi})^{f_{hi2}} / f_{hi1}!\, f_{hi2}!, \tag{3.1}$$

where the $\{\theta_{hi}\}$ denote the respective probabilities of good or excellent response for a randomly selected subject with the hth age status and the ith treatment; $\mathbf{f} = (\mathbf{f}'_{11}, \mathbf{f}'_{12}, \ldots, \mathbf{f}'_{42})'$ denotes the concatenated vector of all frequencies in Table 5; and $\boldsymbol{\theta} = (\theta_{11}, \theta_{12}, \ldots, \theta_{42})'$ is a similarly-arranged vector of the $\{\theta_{hi}\}$. One model of interest for the $\{\theta_{hi}\}$ is the linear logistic which can be expressed as

$$\theta_{hi} = [1 + \exp\{-\lambda_i - \xi_i(x_h - 20)/10\}]^{-1}; \tag{3.2}$$

the λ_i are intercept parameters for the respective treatment groups corresponding to subjects of 20 years of age (as a minimum value for the population at which the study was directed); the ξ_i are the slope parameters for the extent to which the logit transformations

$$\eta_{hi} = \ln\{\theta_{hi}/(1-\theta_{hi})\} = \lambda_i + \xi_i(x_h - 20)/10 \tag{3.3}$$

increase as age increases in 10-year units; and the x_h are 'designated midpoint values' of 40, 50, 60 and 70 for the age ranges ≤44, 45–54, 55–64, and ≥65. The linear logistic

Table 5
Summary of logistic model analysis

Age (yrs)	Treatment	Number of patients with response status		Percentage of patients with good or excellent response status			
		Good or excellent	Moderate, fair, poor	Observed	SE	Model ($\mathbf{X}_R$) ML predicted	SE
≤44	Active	1	4	20.0	17.9	39.0	15.7
	Placebo	1	7	12.5	11.7	10.1	6.2
45–54	Active	1	1	50.0	35.4	50.4	11.5
	Placebo	1	8	11.1	10.5	15.1	6.5
55–64	Active	11	2	84.6	10.0	61.7	9.7
	Placebo	2	7	22.2	13.9	22.0	8.0
≥65	Active	3	4	42.9	18.7	71.8	11.0
	Placebo	2	4	33.3	19.2	30.9	12.8

model has been used extensively in the biological sciences; e.g. dose–response relationships in quantal bioassay experiments (see Berkson, 1953; Finney, 1971), paired comparison studies involving Bradley–Terry models (see Bradley, 1976), and epidemiologic investigations concerned with the relationship between a set of risk factors and certain health outcomes (see Anderson *et al.*, 1980; Breslow and Day, 1980; Kleinbaum *et al.*, 1981). For such applications, it can be interpreted as expressing a linear relationship between the log-odds η_{hi} of the two possible outcomes of a binary response, such as good or excellent vs otherwise, and a set of explanatory variables like age and treatment. In this sense, the logistic model (3.2) has the same structure as the type of linear model which is typically used for continuous normally-distributed data to evaluate parallelism (e.g. equality of slopes) as a preliminary to covariance analysis. It also has the advantageous mathematical property that all values of the λ_i and/or ξ_i in $(-\infty, \infty)$ yield values for the $\{\theta_{hi}\}$ in $(0, 1)$.

If most of the frequencies f_{hij} were sufficiently large (e.g. ≥ 5) to have approximately normal distributions, then either weighted least squares (WLS) methods or maximum likelihood (ML) methods could be equivalently used to obtain estimates for the parameters and test statistics for hypotheses of interest. Background references for WLS include Grizzle, Starmer and Koch (1969) and Koch *et al.* (1977); those for ML include Andersen (1980), Bishop, Fienberg, and Holland (1975), Cox (1970), Goodman (1978), Haberman (1978) and Nelder and Wedderburn (1972); Imrey, Koch and Stokes (1981, 1982) review both approaches and discuss certain hybrid combinations. For the example here, many of the $\{f_{hij}\}$ are small, and so ML methods are preferable. These estimates can be expressed as the solution of the nonlinear equations derived by substituting the model expression (3.2) for the $\{\theta_{hi}\}$ into the expression (3.1) for ϕ, differentiating $\ln \phi$ with respect to the λ_i and ξ_i and equating the result to 0. The equations obtained by these steps are

$$\sum_{h=1}^{4} n_{hi}\hat{\theta}_{hi} = \sum_{h=1}^{4} f_{hi1} = L_{i1}, \qquad \sum_{h=1}^{4} h n_{hi}\hat{\theta}_{hi} = \sum_{h=1}^{4} h f_{hi1} = L_{i2}, \tag{3.4}$$

where the $\hat{\theta}_{hi} = [1 + \exp\{-\hat{\lambda}_i - \hat{\xi}_i(x_h - 20)/10\}]^{-1}$ are the model-predicted estimates of the θ_{hi} based on the ML parameter estimates $\hat{\lambda}_i$ and $\hat{\xi}_i$. Since the equations (3.4) are nonlinear, iterative procedures are required to compute the $\hat{\lambda}_i$ and $\hat{\xi}_i$. For this purpose, one useful approach is the Newton–Raphson method (or iterative weighted least squares). This type of computing procedure yielded the estimates and estimated standard errors shown in the upper part of Table 6 with respect to the matrix formulation $\mathbf{logit}(\boldsymbol{\theta}) = \mathbf{X}\boldsymbol{\beta}$ for the model where $\mathbf{logit}(\boldsymbol{\theta})$ is the vector of $\ln\{\theta_{hi}/(1-\theta_{hi})\}$-transformed values of the θ_{hi}, $\boldsymbol{\beta} = (\lambda_1, \xi_1, \lambda_2, \xi_2)'$ is the vector of parameters, and $\mathbf{X}$ is the model-specification matrix for the structure (3.3) for the respective age × treatment subpopulations. In this setting, the estimated standard errors are obtained from the estimated covariance matrix

$$\mathbf{V}_{\hat{\boldsymbol{\beta}}} = \left\{ \sum_{h=1}^{4} \sum_{i=1}^{2} n_{hi}\hat{\theta}_{hi}(1-\hat{\theta}_{hi})\mathbf{x}_{hi}\mathbf{x}'_{hi} \right\}^{-1}, \tag{3.5}$$

where $\mathbf{x}'_{hi}$ is the row of $\mathbf{X}$ for the (h, i)th group.

The goodness of fit of the model (3.2) can be assessed by using either the log likelihood ratio chi square statistic

$$Q_L = \sum_{h=1}^{4} \sum_{i=1}^{2} 2(f_{hi1}\{\ln(f_{hi1}/n_{hi}\hat{\theta}_{hi})\} + f_{hi2}[\ln\{f_{hi2}/n_{hi}(1-\hat{\theta}_{hi})\}]) = 6.89, \tag{3.6}$$

Table 6
Estimated parameters for logistic models for good or excellent response

Preliminary logistic model design matrix and ML estimates for parameters

Specification matrix, $\mathbf{X}$	Parameter interpretation	ML logistic model parameter estimates ±SE
$\begin{bmatrix}1&2&0&0\\0&0&1&2\\1&3&0&0\\0&0&1&3\\1&4&0&0\\0&0&1&4\\1&5&0&0\\0&0&1&5\end{bmatrix}$	$\begin{bmatrix}\lambda_1\text{: Intercept for active treatment at age 20}\\ \xi_1\text{: Linear increment for (age-20)/10 for active treatment}\\ \lambda_2\text{: Intercept for placebo treatment at age 20}\\ \xi_2\text{: Linear increment for (age-20)/10 for placebo treatment}\end{bmatrix}$	$\begin{bmatrix}-1.30\pm1.54\\ 0.44\pm0.40\\ -3.20\pm1.75\\ 0.49\pm0.45\end{bmatrix}$

Final logistic model design matrix and ML estimates for parameters

Specification matrix, $\mathbf{X}_R$	Parameter interpretation	ML logistic model parameter estimates ±SE
$\begin{bmatrix}1&2&0\\1&2&1\\1&3&0\\1&3&1\\1&4&0\\1&4&1\\1&5&0\\1&5&1\end{bmatrix}$	$\begin{bmatrix}\lambda_{1R}\text{: Intercept for active treatment at age 20}\\ \xi_R\text{: Linear increment for (age-20)/10 for all patients}\\ \tau_R\text{: Increment for placebo treatment}\end{bmatrix}$	$\begin{bmatrix}-1.37\pm1.19\\ 0.46\pm0.30\\ -1.74\pm0.61\end{bmatrix}$

or the Pearson chi square criterion

$$Q_P = \sum_{h=1}^{4}\sum_{i=1}^{2}\{(f_{hi1}-n_{hi}\hat{\theta}_{hi})^2/n_{hi}\hat{\theta}_{hi}(1-\hat{\theta}_{hi})\} = 6.70, \tag{3.7}$$

each of which has $s-u=4$ df, where $u=4$ is the rank of $\mathbf{X}$. Although both Q_L and Q_P are nonsignificant with $P>.10$, they need to be interpreted with caution because the presence of many small frequencies $\{f_{hij}\}$ in Table 5 tends to contradict the strict validity of chi square approximations. Nevertheless, numerical studies such as Larntz (1978) suggest that chi square approximations are reasonable for evaluating goodness of fit for a broad range of small-sample situations, particularly with Q_P when most of the model-predicted frequencies exceed 2 and few are less than 1. Thus, the variation among the $\hat{\theta}_{hi}$ is judged here to be compatible with the model (3.2).

Since the $\hat{\lambda}_i$ and $\hat{\xi}_i$ are implicitly functions of the linear statistics L_{i1} and L_{i2} for which arguments based on the central limit theory and Taylor series are applicable, they have approximately a multivariate normal distribution. Linear hypotheses of the type $\mathbf{C}\boldsymbol{\beta}=\mathbf{0}$, where $\mathbf{C}$ is a specified matrix of full-rank $c\leq u=4$, can therefore be tested via the

generalized Wald criterion

$$Q_{WC} = \hat{\boldsymbol{\beta}}\mathbf{C}'(\mathbf{C}\mathbf{V}_{\hat{\boldsymbol{\beta}}}\mathbf{C}')^{-1}\mathbf{C}\hat{\boldsymbol{\beta}} \tag{3.8}$$

which has an approximate χ^2 distribution with c df under the hypothesis. Alternatively, one can use the log likelihood reduction which is minus twice the difference between the log likelihood statistic (3.6) with respect to $\hat{\beta}$ for $\mathbf{X}$ and its counterpart with respect to $\hat{\boldsymbol{\gamma}}$ for the model $\mathbf{X}_R = \mathbf{XZ}$ implied by the hypothesis (via $\boldsymbol{\beta} = \mathbf{Z}\boldsymbol{\gamma}$ where $\mathbf{Z}$ is an orthocomplement to $\mathbf{C}$). For further discussion, see Bishop *et al.* (1975) and Imrey *et al.* (1981, 1982).

For the model $\mathbf{X}$, three hypotheses are of interest. One is the hypothesis, $\xi_1 - \xi_2 = 0$, of parallelism of the linear logistic relationships (3.3) for the two treatment groups. Since its test statistic $Q_{WC} = 0.01$ with 1 df was nonsignificant with $P > .25$, the model (3.2) can be simplified to have a common slope parameter ξ_R. The corresponding specification matrix $\mathbf{X}_R$ is shown in the lower part of Table 6. The other two hypotheses concerning the model $\mathbf{X}$ parameters were directed at the equality of the average treatment difference for two definitions of population average age; these were the hypotheses $(\lambda_1 - \lambda_2) + 3.5(\xi_1 - \xi_2) = 0$ for the middle value of 55 for the set of age intervals $h = 1, 2, 3, 4$, and $(\lambda_1 - \lambda_2) + 3.35(\xi_1 - \xi_2) = 0$ for the average value of 53.5 for the overall set of 59 patients in the study. Since $Q_{WC} = 7.73$ (for the former) and $Q_{WC} = 7.16$ (for the latter) are both significant with $P < .01$, the active treatment can be interpreted as being more effective, on average, than placebo for such populations. Although this type of conclusion is of definite interest because it is based upon the adjusted comparison of treatments at the same age value, it does not necessarily imply that the active treatment is better at every age. However, as noted previously, the $\{\theta_{hi}\}$ are compatible with the parallel linear logistic model $\mathbf{X}_R$ in the sense that both the goodness-of-fit statistics $Q_L = 6.90$ and $Q_P = 6.74$ with 5 df are nonsignificant with $P > .10$. Accordingly, from the corresponding estimated parameters $\hat{\boldsymbol{\gamma}}$ and their estimated standard errors, it can be verified that the test statistic $Q_{WC} = (1.74/0.61)^2 = 8.09$ with 1 df for comparing treatments was significant with $P < .01$. Thus, the analysis of this model provides the basis for the generalizability of the conclusion that active treatment was more effective than placebo to the entire age range for the population of which the subjects in the study are considered representative.

The estimated parameters $\hat{\boldsymbol{\gamma}} = (\hat{\lambda}_{1R}, \hat{\xi}_R, \hat{\tau}_R)'$ in Table 6 for the model $\mathbf{X}_R$ can be used to construct predicted values $\hat{\boldsymbol{\theta}}_R$ for the $\{\theta_{hi}\}$ and their estimated covariance structure $V_{\hat{\boldsymbol{\theta}},R}$, using matrix computational procedures like those in Koch *et al.* (1977, Appendix 1) and Imrey *et al.* (1981, 1982). The predicted values $\hat{\boldsymbol{\theta}}_R$ are given on the right-hand side of Table 5 together with their estimated standard errors. These quantities describe the extent to which the probability of good or excellent response is greater for active treatment than placebo for each specific age group and its increasing relationship with age for each treatment; they also have smaller standard errors than the observed proportions $\{f_{hi1}/n_{hi}\}$.

Predicted values $\hat{\theta}_{R,i}(x)$ can also be obtained for any specific value of age x_0 such as $x_0 = 53.5$ for the average age of all 59 patients in the study. These quantities can be interpreted as 'adjusted means'. Their resulting values (and estimated standard errors) are $\hat{\theta}_{R,1}(53.5) = 0.544(0.104)$ and $\hat{\theta}_{R,2}(53.5) = 0.173(0.068)$. Since the unadjusted proportions of good or excellent responses from Table 2 were $\bar{a}_1 = \frac{16}{27} = 0.593$ and $\bar{a}_2 = \frac{6}{32} = 0.188$, with estimated standard errors 0.095 and 0.069 respectively, the parallel-line logistic model covariance analysis can be seen to lead to 'adjusted means' which are nearer together than their unadjusted counterparts, as was also the case for the randomization covariance analysis in §2.2. Another similarity between the results of these two methods is the extent to which their test statistics for the comparison of the active treatments are not as strongly significant as the corresponding direct test from §2.1 (i.e. both $Q_{RC} = 8.79$ from Table 3

and $Q_{WC} = 8.08$ from the model $\mathbf{X}_R$ in Table 6 are smaller than $Q_{RS} = 10.10$ in Table 3, each with 1 df). However, the rationale for logistic-model covariance analysis is the same as that for the randomization analysis in §2.2; both provide more explicit conclusions concerning treatment effects via their respective frameworks for adjustment for subject heterogeneity with respect to covariables. The results of the logistic-model analysis have the additional advantage of applying to patients of any age in the conceptual population represented here. Its scope is more comprehensive than that of the randomization methods, but this is achieved via external assumptions about the sampling process for the data as the basis for the underlying product-binomial distribution (3.1). In other words, neither approach is preferred because the assumptions and objectives are different. Further discussion of such considerations and related issues is given in Koch *et al.* (1980).

In the previous discussion, attention was directed at the frequencies in Table 5 which expressed age as four age ranges. Since the $\{\theta_{hi}\}$ were compatible with the linear model (3.2), it is of interest to investigate whether such relationships also apply to the individual responses of the respective subjects. For this purpose, let $a_{il} = 1$ if the lth subject in the ith group has good or excellent response and let $a_{il} = 0$ otherwise; let x_{i*l} denote that subject's actual age. Also, let $\theta_i(x_{i*l}) = \mathrm{E}\{a_{il}\}$ denote the probability of good or excellent response for subjects with age x_{i*l}. Then the model for individual responses which is analogous to (3.2) is

$$\theta_i(x_{i*l}) = [1 + \exp\{-\lambda_i^* - \xi_i^*(x_{i*l} - 20)/10\}]^{-1}, \tag{3.9}$$

and its parallel-line counterpart analogous to $\mathbf{X}_R$ is

$$\theta_i(x_{i*l}) = [1 + \exp\{-\lambda_{1R}^* - \xi_R^*(x_{i*l} - 20)/10\} + \tau_R^* w_i]^{-1}, \tag{3.10}$$

where $w_1 = 0$, $w_2 = 1$ is a placebo–treatment indicator variable. Maximum likelihood estimates for the parameters of the models (3.9) or (3.10) and their estimated covariance structure can be obtained by the same types of computational procedures as described for the model (3.2). In particular, for the model (3.10), the computer program LOGIST documented in Harrell (1980) was used to obtain the estimates and their estimated standard errors shown in (3.11).

	λ_{1R}^*	ξ_R^*	τ_R^*
MI estimate	−1.07	0.41	−1.77
Estimated SE	1.02	0.26	0.61

(3.11)

Since the estimates from this raw-data framework are very similar to those for the corresponding contingency table framework as given for the model $\mathbf{X}_R$ in the lower part of Table 6, the loss of information due to grouping can be interpreted as being minimal for the data in Table 1. Thus, it is possible to fit linear-logistic models to raw data like Table 1 as well as contingency tables. The former is often preferable when there is a substantive basis for the model (3.9) or (3.10), the covariables x_{i*l} are measured without error, and the sample size is moderate (e.g. $\leq$500) with respect to computational load; otherwise, the latter may be a more practical approach for dealing with goodness-of-fit issues when an *a priori* model is not available and for cost savings in the analysis of large data sets (e.g. $\geq$2000 observations).

Finally, if the data in Table 5 had not been compatible with the linear logistic model (3.2), then attention would have been directed at the model

$$\theta_{hi} = \{1 + \exp(-\lambda - \tau_i - \delta_h)\}^{-1}, \tag{3.12}$$

where λ is a reference cell parameter for active-treatment patients with age ≤ 44, $\tau_1 = 0$ and τ_2 are treatment parameters, and $\delta_1 = 0$, δ_2, δ_3 and δ_4 are age-group parameters. In this model, there are three covariables for age which are expressed as three indicator variables rather than a single covariable for age-range midpoints. Since the goodness-of-fit tests $Q_L = 3.21$ and $Q_P = 3.27$ with 3 df are both nonsignificant for the model (3.13), further analysis of its parameters is warranted. The ML estimate for the treatment parameter was $\hat{\tau}_2 = -1.70$, and the corresponding Wald statistic for comparing it to 0 was $Q_{WC} = 6.90$ ($P < .01$). Thus, the test statistic for treatment effects from the model (3.12) is similar to its randomization-model counterparts in Table 3 for which the age groups were used either as multiple indicator covariables or as strata. Otherwise, if the number of patients in each group is small, conditional likelihood methods may be needed to assess treatment effects in a setting like (3.12); see Breslow and Day (1980), Holford (1982) or Kleinbaum, Kupper and Chambless (1982).

3.2 *Log-Linear Model Analysis for Entire Response Distribution*

The logistic-model analysis discussed in §2.1 for the binary outcome of good or excellent response can be extended to the $r = 5$ category distributions of patient response status for the $s = 8$ subgroups shown in Table 4 by the use of log-linear models. In this case, the frequency vectors $\{\mathbf{n}_{hi} = (n_{hi1}, n_{hi2}, n_{hi3}, n_{hi4}, n_{hi5})'\}$ are assumed to have the product-multinomial distribution

$$\phi(\mathbf{n} \mid \boldsymbol{\pi}) = \prod_{h=1}^{4} \prod_{i=1}^{2} n_{hi}! \left\{ \prod_{j=1}^{5} (\pi_{hij}^{n_{hij}} / n_{hij}!) \right\}, \tag{3.13}$$

where the $\{\pi_{hij}\}$ denote the probabilities of the $j = 1, 2, 3, 4, 5$ response categories for a randomly selected subject with the hth age status and ith treatment from the population which the data conceptually represent, and thereby are subject to the constraints $\sum_{j=1}^{5} \pi_{hij} = 1$ for $h = 1, 2, 3, 4$ and $i = 1, 2$; $\mathbf{n} = (\mathbf{n}'_{11}, \mathbf{n}'_{12}, \ldots, \mathbf{n}'_{42})'$ denotes the concatenated vector of all frequencies in Table 4; and $\boldsymbol{\pi}$ denotes the similarly-arranged vector of the $\{\pi_{hij}\}$. The log-linear model extension of (3.2) to the $r = 5$ categories of patient response status can be expressed as

$$\pi_{hij} = \exp(\psi_{hij}) \Big/ \sum_{j=1}^{5} \exp(\psi_{hij}) \tag{3.14}$$

with $\psi_{hij} = \{\lambda_{ij} + \xi_{ij}(x_h - 20)/10\}$, where the $\{\lambda_{ij}\}$ denote response-category intercept parameters and the $\{\xi_{ij}\}$ denote slope parameters on age for each treatment group; also to avoid redundancies in the model specification, its parameters are defined so that $\lambda_{i5} = \xi_{i5} = 0$ for $i = 1, 2$. Moreover, the context in which the $\{\lambda_{ij}\}$ are intercept parameters and the $\{\xi_{ij}\}$ are slope parameters follows from the fact that (3.14) implies

$$\ln\{\pi_{hij}/\pi_{hi5}\} = \lambda_{ij} + \xi_{ij}(x_h - 20)/10 \tag{3.15}$$

for $j = 1, 2, 3, 4$ which can be interpreted as an $r - 1 = 4$-variate extension of (3.3). Hence, there are $4 \times 2 = 8$ intercept parameters and $4 \times 2 = 8$ slope parameters in all.

The reduced parallel-line model corresponding to $\mathbf{X}_R$ in the lower part of Table 6 can be expressed in the form (3.14) with

$$\psi_{hij} = \{\lambda_{1jR} + \xi_{jR}(x_h - 20)/10 + \tau_{jR} w_i\}, \tag{3.16}$$

where $w_1 = 0$, $w_2 = 1$ is a placebo–treatment indicator variable, the $\{\tau_{jR}\}$ are treatment-effect parameters, and the $\{\xi_{jR}\}$ are common slope parameters, and the parameters are so defined that $\lambda_{15R} = \tau_{5R} = \xi_{5R} = 0$.

Since most of the frequencies $\{n_{hij}\}$ in Table 4 are small (i.e. ≤ 2), maximum likelihood (ML) methods are used to estimate the parameters. Iterative procedures such as the Newton–Raphson method (or iterative weighted least squares) are usually necessary. Further discussion of the ML estimates and their estimated covariance matrix is given in Andersen (1980), Fienberg (1980), Haberman (1978), and Imrey *et al.* (1981, 1982). For the data in Table 4, the maximum likelihood estimates (and their estimated standard errors) for the model (3.16) are given in (3.17):

$$\begin{array}{llll} \hat{\lambda}_{11R} = -6.27\ (2.85) & \hat{\lambda}_{12R} = -1.31\ (1.43) & \hat{\lambda}_{13R} = -3.90\ (1.79) & \hat{\lambda}_{14R} = -1.46\ (1.77) \\ \hat{\xi}_{1R} = 1.61\ (0.67) & \hat{\xi}_{2R} = 0.61\ (0.39) & \hat{\xi}_{3R} = 1.06\ (0.44) & \hat{\xi}_{4R} = -0.05\ (0.45) \\ \hat{\tau}_{1R} = -1.63\ (1.08) & \hat{\tau}_{2R} = -1.81\ (0.81) & \hat{\tau}_{3R} = -0.40\ (0.85) & \hat{\tau}_{4R} = 1.07\ (1.19) \end{array} \tag{3.17}$$

The goodness of fit of the model (3.16) can be assessed by either the log likelihood ratio χ^2 statistic $Q_L = 19.62$ or the Pearson χ^2 criterion $Q_P = 17.59$, both of which are nonsignificant with $P \geq 0.25$ relative to χ^2 approximations to their distributions [with $s(r-1)-u = 20$ df, where $u = 12$ is the rank of $\mathbf{X}$ for the model (3.16)]. Further analyses of the estimated parameters (3.17) in either of two directions are warranted. One is to use Wald statistics like (3.8) to test hypotheses about specific sources of variation such as age and treatment. Here, the approximate normality of the estimated parameters follows from their being implicit functions of linear statistics analogous to (3.4) to which arguments based on the central limit theory are considered applicable (Imrey *et al.*, 1981, 1982). For the hypothesis $\xi_{1R} = \xi_{2R} = \xi_{3R} = \xi_{4R} = 0$ of no age effects, $Q_{WC} = 10.05$ with 4 df is significant ($P = .040$); and for the hypothesis $\tau_{1R} = \tau_{2R} = \tau_{3R} = \tau_{4R} = 0$ of no treatment effects, $Q_{WC} = 8.84$ with 4 df is suggestive ($P = .065$), but the log likelihood ratio χ^2 statistic $Q_{LC} = 30.11 - 19.62 = 10.49$ is significant ($P = .032$). Thus, the latter results are similar to their randomization-test counterparts in Table 3, but weaker than the corresponding direct test.

The other types of hypotheses which are of interest for the model (3.16) are concerned with its simplification to a final model for which information in the ordinal scale for the response variable is taken into account. This can be achieved by identifying a mechanism such that the age effects and treatment effects can be expressed in terms of single parameters rather than sets of four parameters. One strategy of this type is to require that the within-age-group, adjacent-response-category ln(odds ratios) for treatments,

$$\ln(\pi_{h1j}\pi_{h2,(j+1)}/\pi_{h1,(j+1)}\pi_{h2j}) = \tau_{(j+1),R} - \tau_{jR}, \tag{3.18}$$

be equal for $j = 1, 2, 3, 4$; and the within-treatment group, adjacent-response-category ln(odds ratios) for adjacent ages,

$$\ln(\pi_{hij}\pi_{(h+1),i,(j+1)}/\pi_{hi,(j+1)}\pi_{(h+1),ij}) = \xi_{(j+1),R} - \xi_{jR}, \tag{3.19}$$

be equal for $j = 1, 2, 3, 4$. Since the parameters $\{\xi_{jR}\}$ and $\{\tau_{jR}\}$ are defined with $\tau_{5R} = \xi_{5R} = 0$ to avoid redundancies, the hypotheses corresponding to (3.18) and (3.19) are $4\tau_{1R} = 3\tau_{2R} = 2\tau_{3R} = \tau_{4R}$ and $4\xi_{1R} = 3\xi_{2R} = 2\xi_{3R} = \xi_{4R}$, respectively. The corresponding test statistics are $Q_{WC} = 2.96$ with 3 df for simplification of treatment effects, $Q_{WC} = 4.11$ with 3 df for simplification of age effects, and $Q_{WC} = 6.90$ with 6 df for simultaneous simplification of age and treatment effects. Since all these test statistics are nonsignificant with $P \geq .25$, both types of simplifications are considered warranted. Accordingly, the resulting model can be expressed in the form (3.14) with

$$\psi_{hij} = \{\lambda_{1jRR} + \xi_{*RR}(5-j)(x_h - 20)/10 + \tau_{*RR}(5-j)w_i\}; \tag{3.20}$$

here, the $\{\lambda_{1jRR}\}$ are response-category intercept parameters for the active treatment, ξ_{*RR} is a common age-slope parameter for all treatment × response-category combinations, and τ_{*R} is a common placebo–treatment effect for all age group × response combinations. The maximum likelihood estimates for the parameters of the model (3.20) and their estimated standard errors are given in (3.21).

	λ_{11RR}	λ_{12RR}	λ_{13RR}	λ_{14RR}	$\xi_{*_{RR}}$	τ_{*RR}
ML estimate	−4.26	−2.39	−1.67	−1.32	0.31	−0.58
Estimated SE	1.79	1.29	0.87	0.57	0.11	0.23

(3.21)

Predicted values for the π_{hij} and their estimated standard errors can be obtained by matrix computations like those in Imrey *et al.* (1981, 1982). Such quantities corresponding to 'adjusted means' at the average age of 53.5 for all patients in the study are given in Table 7. Since the goodness-of-fit test statistics $Q_L = 27.70$ and $Q_P = 24.43$, each with 26 df, are nonsignificant ($P \geq .25$), the variation among the $\{\pi_{hij}\}$ is concluded to be compatible with the model (3.20). Thus, the hypothesis of no treatment effects, $\tau_{*RR} = 0$, can be tested via the Wald statistic $Q_{WC} = \{(-0.58/0.23)^2\} = 6.27$ which is significant with $P = .012$ from its approximate χ^2 distribution with 1 df. This result is of additional interest because it is the log-linear model counterpart of the direct and 'actual-age' adjusted randomization tests in Table 3 based on integer scores, relative to which it is somewhat smaller. In fact, it can be shown that the estimated parameters (3.21) are implicitly linear functions of the overall numbers of subjects $n_{++j} = \sum_{h=1}^{4}\sum_{i=1}^{2} n_{hij}$ with the jth response level and the linear functions $\sum_{h=1}^{4}\sum_{i=1}^{2}(h+1)\tilde{a}_{hi}$ and $\sum_{h=1}^{4}\sum_{i=1}^{2}(i-1)\tilde{a}_{hi}$ of the mean scores $\tilde{a}_{hi} = \sum_{j=1}^{5}(5-j)n_{hij}$ for the (h, i)th group, with the integer scores (4, 3, 2, 1, 0). This aspect and other issues concerning the use of log-linear models like (3.20) for ordinally-scaled data have been previously discussed by Haberman (1974), Andrich (1979), Goodman (1979), Imrey *et al.* (1981, 1982), and elsewhere. In particular, Andrich (1979) provides a psychometric rationale for why model simplifications like those pertaining to (3.18) and (3.19) are plausible for ordinal data provided that the $\{\lambda_{1jRR}\}$ increase for $j = 1, 2, 3, 4$ (which does hold for their estimates here). On the other hand, McCullagh (1980) indicates that other models for ordinal data, expressed in terms of the cumulative probabilities $\phi_{hij} = \sum_{k=1}^{j}\pi_{hik}$, may be preferable to log-linear models like (3.16) or (3.20) for many situations. Other references dealing with alternative methods for the analysis of ordinal data in a cross-classified data setting include Agresti (1980), Clogg (1982), Semenya and Koch (1979, 1981), and Simon (1974).

Table 7
Maximum likelihood estimates (and standard errors) for log-linear model at average age of 53.5 *years*

Treatment	Global evaluation of patient at end of study				
	Excellent	Good	Moderate	Fair	Poor
Active	0.146	0.333	0.241	0.120	0.159
	(0.060)	(0.079)	(0.062)	(0.045)	(0.067)
Placebo	0.038	0.156	0.202	0.180	0.424
	(0.023)	(0.052)	(0.055)	(0.058)	(0.087)

4. Weighted Linear Model Analyses for Correlated Marginal Means

Another conceptual sampling framework for the data in Table 1 is to presume that the subjects under study for each treatment are representative of some large population for the bivariate distribution of response status and age. In other words, the procedures by which subjects are recruited to participate in the study do not involve any selection bias with respect to age, so subjects with certain ages are neither more nor less likely to be included in the study than those with other ages. This assumption is not always realistic in practice, and so the results may not be meaningful for the target population at which a study is directed. As in §3, it is assumed that the response of each subject is representative of the conditional distribution of subjects with the same age.

Comparisons between treatment groups can be undertaken by fitting linear models via weighted least squares (WLS) to the linear sample statistics $\mathbf{F} = (\bar{\mathbf{y}}_1', \bar{\mathbf{x}}_1', \bar{\mathbf{y}}_2', \bar{\mathbf{x}}_2')'$ where the $\{(\bar{\mathbf{y}}_i', \bar{\mathbf{x}}_i')'\}$ are within-group mean vectors as in (2.8). However, for the large-population (or with-replacement) sampling framework presumed here, the estimated covariance matrix $\mathbf{V}_F$ for $\mathbf{F}$ is a block diagonal matrix of within-treatment group covariance matrices

$$\mathbf{V}_i = \frac{1}{n_i^2} \sum_{l=1}^{n_i} \left(\begin{bmatrix} (\mathbf{y}_{il} - \bar{\mathbf{y}}_i) \\ (\mathbf{x}_{il} - \bar{\mathbf{x}}_i) \end{bmatrix} [(\mathbf{y}_{il} - \bar{\mathbf{y}}_i)', (\mathbf{x}_{il} - \bar{\mathbf{x}}_i)'] \right), \tag{4.1}$$

as opposed to being based on (2.9). Two sets of functions are considered,

$$\begin{aligned} \mathbf{F}_1 &= (\bar{a}_1, \bar{x}_{1*}, \bar{a}_2, \bar{x}_{2*})', \\ \mathbf{F}_2 &= (\mathbf{p}_1', \bar{x}_{1*}, \mathbf{p}_2', \bar{x}_{2*})', \end{aligned} \tag{4.2}$$

where the $\{\bar{a}_i\}$ are mean values of response status for the two treatment groups with respect to integer scores, the $\{\mathbf{p}_i\}$ are the mean vectors of indicator variables yielding the proportions $p_{ij} = n_{ij}/n_{i+}$ of subjects with the respective response-status outcomes $j = 1, 2, 3, 4, 5$ and the $\bar{x}_{i*}$ are the mean ages. The observed values of the functions $\mathbf{F}_1$ and their estimated standard errors are shown in the third and fourth columns of Table 8, while those for $\mathbf{F}_2$ are shown in the third and fourth columns of Table 9.

Since the sample sizes $\{n_{+i} = 27, 32\}$ for this example are considered sufficiently large for mean vectors like $\mathbf{F}_1$ and $\mathbf{F}_2$ to have multivariate normal distributions, the variation among their elements can be investigated by the weighted least squares methods discussed by Grizzle *et al.* (1969) and Koch *et al.* (1977). Thus, covariance adjustment can be undertaken by fitting linear models which involve the constraint that no difference is expected in the mean ages of the two treatment groups. For the function vector $\mathbf{F}_1$, this

Table 8
Weighted least squares analysis of patient response with covariance adjustment for age

Treatment group	Summary function	Observed values		Covariance model: WLS predicted values	
		Estimate	SE	Estimate	SE
Active	Mean response status	2.63	0.26	2.72	0.25
	Mean age	55.74	2.28	53.71	1.62
Placebo	Mean response status	3.72	0.22	3.66	0.22
	Mean age	51.62	2.31	53.71	1.62

Table 9
Weighted least squares analysis with covariance-oriented multivariate linear model

Treatment group	Summary functions for patient response status and age	Within-treatment-group observed mean values		Covariance-oriented multivariate model: WLS predicted values	
		Estimate	SE	Estimate	SE
Active	% excellent	18.52	7.48	24.00	5.48
	% good	40.74	9.46	26.87	5.04
	% moderate	18.52	7.48	21.45	5.19
	% fair	3.70	3.63	6.00	3.32
	% poor	18.52	7.48	21.68	6.33
	Mean age	55.74	2.28	53.93	1.61
Placebo	% excellent	6.25	4.28	5.74	4.07
	% good	12.50	5.85	17.75	5.16
	% moderate	21.88	7.31	21.45	5.19
	% fair	21.88	7.31	15.13	3.87
	% poor	37.50	8.56	39.94	5.96
	Mean age	51.63	2.31	53.93	1.61

model can be expressed as

$$\mathrm{E}(\mathbf{F}_1) = \mathbf{X}_S\boldsymbol{\mu} = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \\ 0 & 1 & 0 \end{bmatrix} \begin{bmatrix} \mu_{11} \\ \mu_{*2} \\ \mu_{21} \end{bmatrix}, \tag{4.3}$$

where μ_{11} and μ_{21} are 'adjusted mean value' parameters for patient response status for the two treatment groups and μ_{*2} is the common mean-age parameter. The estimated parameters $\hat{\boldsymbol{\mu}}$ from the WLS fit of the model $\mathbf{X}_S$ and their estimated covariance matrix $\mathbf{V}_{\hat{\mu}}$ are obtained via the computations

$$\hat{\boldsymbol{\mu}}' = (\mathbf{X}_S'\mathbf{V}_{\mathbf{F}_1}^{-1}\mathbf{X}_S)^{-1}\mathbf{X}_S'\mathbf{V}_{\mathbf{F}_1}^{-1}\mathbf{F}_1, \qquad \mathbf{V}_{\hat{\mu}} = (\mathbf{X}_S'\mathbf{V}_{\mathbf{F}_1}^{-1}\mathbf{X}_S)^{-1}. \tag{4.4}$$

The resulting estimates (and their estimated standard errors) are $\hat{\mu}_{11} = 2.72\ (0.25)$, $\hat{\mu}_{*2} = 53.71\ (1.62)$, and $\hat{\mu}_{21} = 3.66\ (0.22)$. Since the function vector $\mathbf{F}_1$ plausibly has an approximate multivariate normal distribution with known covariance matrix $\mathbf{V}_{\mathbf{F}_1}$, it follows that $\hat{\boldsymbol{\mu}}$ approximately has a multivariate normal distribution. The goodness of fit of the model $\mathbf{X}_S$ can be assessed by the Wald residual statistic

$$Q_W = (\mathbf{F}_1 - \mathbf{X}_S\hat{\boldsymbol{\mu}})'\mathbf{V}_{\mathbf{F}_1}^{-1}(\mathbf{F}_1 - \mathbf{X}_S\hat{\boldsymbol{\mu}}) = 1.61, \tag{4.5}$$

which is nonsignificant with $P \geq .10$ from its approximate χ^2 distribution with $s(d+t)-u = 1$ df, where s is the number of groups, d is the dimension of the $\mathbf{y}_{il}$, t is the dimension of the $\mathbf{x}_{il}$, and u is the rank of the model $\mathbf{X}_s$. Thus, the variation among the elements of $\mathbf{F}_1$ is concluded to be compatible with the model $\mathbf{X}_S$. Tests of hypotheses concerning $\boldsymbol{\mu}$ can be undertaken via Wald statistics like (3.8). One hypothesis is the comparison $\mu_{11} - \mu_{21} = 0$ of the adjusted treatment means for which $Q_W = 8.68$ with 1 df is significant ($P = .003$). This result is stronger than its randomization-model counterpart in Table 3, but it is weaker than the Wald statistic for comparing the unadjusted means for the sampling

framework here, which is

$$Q_W = (2.63 - 3.72)^2/\{(0.26)^2 + (0.22)^2\} = 10.24 \tag{4.6}$$

with 1 df. The reason why the statistics from the analysis in this section are larger than their counterparts in §2 is that they are based upon within-group covariance matrices like (4.1) in contrast to the pooled-group covariance matrix defined in (2.9). When there is no difference between the treatments, these two covariance matrices would be expected to be similar; but if there is a location-shift difference between them with excellent or good response being more likely in one group, then the within-group variances for estimates of mean response will be smaller than the pooled-group variances; and thereby, the Wald statistics discussed in this section would be expected to be larger than the corresponding statistics in §2. However, this discussion should be interpreted carefully because large-sample χ^2 approximations are used to assess significance in both cases, and these may tend to yield P-values which are too small (relative to what would be obtained by simulations). For this reason, the distinction between the usage of the tests in §2 and those in this section is primarily a matter of the sampling framework, with each being valid in its corresponding setting.

Predicted values $\mathbf{F}_1 = \mathbf{X}_S\hat{\boldsymbol{\mu}}$ from the model $\mathbf{X}_S$ are shown in the upper right part of Table 8 together with their estimated standard errors, as obtained from square roots of diagonal elements of $\mathbf{X}_S\mathbf{V}_{\hat{\boldsymbol{\mu}}}\mathbf{X}_S'$. These quantities can be seen to be nearer together than their unadjusted counterparts as was also the case for those from the methods discussed in §2 and §3. Two other aspects of the predicted values $\hat{\mathbf{F}}_1$ and their standard errors are of note. First, they can be used to construct confidence intervals for the mean response (for integer scores) and for the difference between mean responses. This capability also applies to the predicted values from the models in §3, but not to the adjusted means from §2 since the covariance matrix (2.9) for the latter is linked to hypotheses like H_{03} in (2.7) of no association between treatment and the joint distributions of age and response status. Secondly, the use of the model $\mathbf{X}_S$ and its predicted values requires only that the mean ages for the two groups are sufficiently similar that the goodness-of-fit test (4.5) is nonsignificant. Tests of parallelism are not needed, as was the case in §3, because no predictive model with respect to age is involved. However, the methods in this section yield conclusions only about adjusted means and not about the relationship between response and age for the two treatment groups, and so they may not necessarily be generalizable to all specific values of age.

Two additional types of WLS analysis of marginal means are illustrative. One is covariance adjustment for the entire response distributions via the functions $\mathbf{F}_3 = [\mathbf{F}_{31}', \mathbf{F}_{32}']'$, where

$$\mathbf{F}_{3i} = \begin{bmatrix} 1 & 2 & 3 & 4 & 5 & 0 \\ -1 & 2 & -1 & 0 & 0 & 0 \\ 0 & -1 & 2 & -1 & 0 & 0 \\ 0 & 0 & -1 & 2 & -1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} \mathbf{p}_i \\ \bar{x}_{i*} \end{bmatrix} \tag{4.7}$$

with $\mathbf{p}_i$ and $\bar{x}_{i*}$ being components of the vector $\mathbf{F}_2$ in (4.2). If the difference between the response distributions for the two groups can be characterized by the difference in their mean values $\{\bar{a}_i\}$ with respect to integer scores (i.e. there is no between-group variation

for the other four types of functions in $\mathbf{F}_3$), then the model $\mathbf{X}_{MS}$ in (4.8) will be applicable.

$$\mathrm{E}(\mathbf{F}_3) = \mathbf{X}_{MS}\boldsymbol{\mu}^* = \begin{bmatrix} \mathbf{I}_5 & \mathbf{0}_5 \\ \mathbf{I}_5 & \mathbf{K} \end{bmatrix} \boldsymbol{\mu}^*, \tag{4.8}$$

where $\mathbf{K} = (1, 0, 0, 0, 0)'$. Since the goodness-of-fit statistic $Q_W = 5.21$ with 4 df is nonsignificant ($P \geq .25$) for the model (4.8), further analysis based on its estimated parameters $\hat{\boldsymbol{\mu}}^* = [2.74, 0.08, 0.10, -0.31, 53.93, 0.91]'$ is warranted. The Wald statistic $Q_W = 9.78$ with 1 df for the hypothesis $\mu_6^* = 0$ of no difference between the adjusted treatment means is significant ($P \leq .01$). Predicted values for the proportions of subjects with the respective response-status outcomes for each treatment group are obtained by matrix operations which invert the transformations $\mathbf{A}_3$ in (4.7) relative to the constraint that such proportions add to 1. These 'adjusted mean distributions', which are given in Table 9, are analogous to those from the log-linear model (3.20) in Table 7; both are from models which express the difference between the two groups in terms of a single parameter which is linked to the mean response status with respect to integer scores. However, the 'adjusted mean distributions' obtained from the two methods are different because one is based on log-linear functions and the other is based on linear functions.

The third application of WLS methods employs the functions

$$U_1 = \sum_{j=1}^{5} p_{1j}\left\{\left(\sum_{k=j+1}^{5} p_{2k}\right) + \tfrac{1}{2}p_{2j}\right\}, \qquad U_2 = \bar{x}_{1*} - \bar{x}_{2*}. \tag{4.9}$$

Here, U_1 is an estimate of the probability that the response of an active-treatment subject is at least as favorable as that of a placebo-treatment subject in the sense of the Mann–Whitney nonparametric test with adjustment for ties, and U_2 is the difference between the mean ages for the two groups. Matrix computational procedures like those in Koch *et al.* (1977, Appendix 1) can be used to obtain the functions $\mathbf{F}_4 = [U_1, U_2]'$ and their estimated covariance matrix $\mathbf{V}_{\mathbf{F}_4} = \mathbf{H}\mathbf{V}_{F_2}\mathbf{H}'$, where $\mathbf{F}_2$ is defined in (4.2) and $\mathbf{H} = \partial\mathbf{F}_4/\partial\mathbf{z}\,|\mathbf{z} = \mathbf{F}_2$ is the first derivative matrix for the linear Taylor series approximation of $\mathbf{F}_4$ relative to $\mathbf{F}_2$. The results of these calculations are

$$\mathbf{F}_4 = \begin{bmatrix} 0.719 \\ 4.12 \end{bmatrix}, \qquad \mathbf{V}_{\mathbf{F}_4} = \begin{bmatrix} 0.004466 & 0.0763 \\ 0.0763 & 10.52 \end{bmatrix}. \tag{4.10}$$

Covariance adjustment is undertaken by fitting the model $\mathbf{X}_{RS} = [1, 0]'$ to $\mathbf{F}_4$ by weighted least squares. Since the goodness-of-fit statistic $Q_W = 1.61$ with 1 df is nonsignificant, the functions $\mathbf{F}_4$ are compatible with the model $\mathbf{X}_{RS}$. The adjusted estimate from this model for the probability that active treatment response is at least as favorable as placebo response is $\hat{U}_1 = 0.689$ with estimated standard error 0.063. Thus; $\hat{U}_1$ is significantly ($P \leq .01$) greater than the value 0.500 which corresponds to treatment equivalence, the Wald statistic for this hypothesis being $Q_W = 9.12$ with 1 df. This result represents the WLS counterpart of the covariance-adjusted randomization tests with respect to midranks in Table 3 except that the latter were adjusted for ranks of age. Otherwise, an important advantage of the WLS estimate $\hat{U}_1$ is that confidence intervals can be obtained from it for the probability that active-treatment response is at least as favorable as placebo response; also, such analysis reflects covariance adjustment for a nonparametric ranking function which takes into account the ordinal nature of the data but involves no scaling assumptions. The applications of WLS methods previously described are illustrative of the broad scope of this approach. Two other examples are given in Koch *et al.* (1978) and Koch and Stokes (1981).

5. Unweighted Linear Model Analyses

In §2, §3 and §4, methods for covariance analysis of categorical data were formulated to account for the measurement scale and the presumed sampling process. Alternatively, it is sometimes practically convenient to apply unweighted least squares to fit covariance-oriented models and to argue that the resulting estimates and test statistics are robust to assumptions of normal distributions and variance homogeneity. For moderate and large samples, this rationale is supported by the conceptual similarity of this heuristic approach to those in §2 and §4 and by the asymptotic normality of the least squares estimates of regression coefficients (for discrete data from bounded intervals). The methods in §2.2 involve the pooled-group covariance matrix in (2.9) for the estimation of adjusted means and residual variance with respect to (2.12); those in §4 involve the within-group covariance matrices in (4.1); and unweighted least squares involves weighted averages of within-group covariance matrices. Thus, these three procedures would be expected to yield similar results for tests of hypotheses that the two treatment groups were equivalent since estimated covariance matrices would relate to essentially the same population structure under the null hypothesis. Table 10 illustrates this point, and provides P-values from F distributions for unweighted least squares covariance analyses of three scalings of patient response status for each of two specifications of covariables; the three scalings were integer scores, midrank scores, and the good or excellent binary indicator; the two covariable specifications were actual age and the age indicators for <44, 45–54 and 55–64. Thus, the test statistics in Table 10 from the unweighted least squares analyses are analogous to those in the third and fourth columns of Table 3; moreover, their corresponding P-values can be seen to be generally similar, although those in Table 10 tend to be smaller. On the other hand, the P-values in Table 10 for integer scores and midrank scores with actual age as the covariable are larger than their weighted least squares counterparts in §4 (see Table 8 and Mann–Whitney function analysis). Thus, the data from this example illustrates the robustness of the unweighted least squares F test for covariance-adjusted comparisons among treatment groups. Further discussion of this general topic for rank transformed data is given in Bennett (1968), Conover and Iman (1981, 1982), and Shirley (1981).

Table 10
General linear model F statistics comparing treatments

General linear model specification	Scaling of patient response status					
	Integer scores		Midrank scores		Good or excellent indicator	
	F	df	F	df	F	df
Actual age; Treatment	8.16 $P = .006$**	(1, 56)	7.72† $P = .007$**	(1, 56)	10.32 $P = .002$**	(1, 56)
Age indicators for <44, 44–54, 55–64; Treatment	4.63 $P = .036$*	(1, 54)	4.72 $P = .034$*	(1, 54)	8.06 $P = .006$**	(1, 54)

* Significant, $.01 < P \leq .05$.
** Significant, $P \leq .01$.
† For this analysis, age corresponds to the ranks of age in the overall sample.

A second type of similarity between the results of unweighted regression analyses and the methods in §2, §3 and §4 is concerned with their respective estimates of adjusted means. For the integer scaling of response, the adjusted means (with estimated standard errors) from covariance analysis on actual age were 2.71 (0.24) for active treatment and 3.65 (0.22) for placebo treatment; while for the binary indicator of good or excellent response, the adjusted proportions were 0.577 (0.086) for active treatment and 0.201

Table 11 *Summary of methodological issues addressed by covariance techniques for categorical data*

Analysis of covariance technique	Sampling framework	Assumptions of linearity parallelism of covariate etc.	Basis for variance estimates	Methods for estimation
Randomization-model covariance-adjusted analysis	No assumption is made that the available sample is a random sample of a larger population; sampling framework is a consequence of randomization	No assumptions made regarding relationship of response to covariables	Hypergeometric distribution	Design-based estimate of difference
Randomization-model stratification-adjusted analysis			Product-hypergeometric distribution	
Log-linear model	It is assumed that the sample available is a random sample of a larger population	Assumptions are necessary and require verification	Product-multinomial distribution	Maximum likelihood
Weighted linear model	Sampling framework can be assumed relative to a larger population or it can be an actual complex sampling framework such as cluster sampling relative to a target population	No assumptions made regarding relationship of response to covariables	Based on presumed or actual sampling framework	Weighted least squares
Unweighted least squares	It is assumed that the sample available is a random sample of a larger population	Assumptions are necessary and require verification	Observed variances and covariances are computed based on assumption of homogeneous variance	Ordinary least squares

(0.079) for placebo treatment. The adjusted means for integer scores from unweighted least squares are similar in value to those from randomization-model methods subsequent to (2.15) and those from weighted least squares in Table 8, and a corresponding similarity of standard errors also holds for the latter. However, the adjusted proportions good or excellent are larger than their counterparts based on the logistic model, $\{\hat{\theta}_{R,i}(53.5)\}$ shown in §3.1, and their standard errors are different with a smaller value for active treatment

Equivalence of distributions of covariates across treatment groups	Conclusions about treatment effects	Generalizability of results	Sample sizes
Needs to be demonstrated (should be a consequence of randomization)	Interpreted in average sense across covariable	Results pertain only to the sample available. Generalization to a larger population must be justified by other arguments.	Exact methods may be feasible. Otherwise small to moderate sample sizes required for asymptotic properties to hold
Not necessarily required as stratification can induce equivalence	Can be specific to level of covariable by considering one stratum. Otherwise, analysis across strata interpreted in average sense		
Not required	Generalizable to all levels of covariable (given parallelism)	Results generalizable to the larger population from which sample is (assumed) drawn	Small–moderate sample sizes for asymptotic properties to hold
Needs to be demonstrated (should be a consequence of randomization)	Interpreted in average sense across covariable	Results generalizable to the larger population from which sample is (assumed) drawn	Moderate sample sizes required for asymptotic properties to hold
Not required	With caution results are generalizable to all levels of covariables (given parallelism)	Results generalizable (with caution) to the larger population from which sample is (assumed) drawn	Moderate–large sample sizes required for asymptotic properties to hold

and a larger one for placebo. One reason for these similarities in adjusted means is that unweighted regression is an unbiased estimation procedure; another is that the within-group estimates of variability tended to be approximately homogeneous for this example. Thus, for other applications, larger differences might occur. Also, for those aspects of analysis which involve predictive models as discussed in §3, the extent to which unweighted least squares provides reasonable estimates of parameters and predicted values is uncertain, particularly for general situations with several cross-classified groups instead of two, more than one covariable, and potential interactions in the group cross-classification, or between groups and covariables. In summary, unweighted least squares can be a useful approximate procedure for the covariance analysis of categorical data, but applications should be undertaken cautiously.

6. Discussion

The natural question arises as to which procedure to use when one is confronted by a specific data-analysis problem. In itself, each covariance technique discussed is relatively straightforward once the mathematical details have been mastered. However, the choice between techniques is not obvious and may be somewhat elusive as some of the pertinent issues are not always elaborated in mathematical discussions of the methods themselves. Unfortunately, a survey paper like the current article is not able to identify the relevant covariance procedure for each practical problem. This would be too ambitious a goal. What can be done, though, is to present a summary of the sorts of issues that may be relevant for an attempt to delineate a particular covariance technique, and to specify the way in which each technique is able to address the relevant issues of analysis. Table 11 has been compiled in this spirit. Columns represent analysis-related issues and rows identify the ways in which each technique deals with these issues. More elaborate discussion of the comments in each cell is provided either in §1 or in the initial statements about each method. In any review of Table 11, the five reasons for undertaking covariance analysis given in §1 should be borne in mind. For example, if generalizability of findings is a major concern, then the fact that randomization-model methods do not allow generalization as a direct consequence of the methods themselves is an important consideration. No single covariance technique is perfect, but a combination of two or more strategies may be robust to a wide variety of considerations.

ACKNOWLEDGEMENTS

The authors would like to thank G. Rex Bryce and the referees for their helpful comments on the original version of this paper. We also would like to express our appreciation to Suzanne Edwards for computational assistance, and JoAnn DeGraffenreidt, Bea Parker, Ann Thomas, and Lori Turnbull for editorial assistance. This research was partially supported by the U.S. Bureau of the Census through Joint Statistical Agreements JSA 79–16 and 80–19.

RÉSUMÉ

On fait un bilan de trois méthodes d'analyse de covariance de données classées en catégories et on les applique à un exemple d'essai clinique sur l'arthrite rhumatismale. Les trois méthodes sont les procédures de randomisation de modèle non paramétrique, la régression logistique par maximum de vraisemblance, l'analyse de moindres carrés pondérés de fonctions marginales corrélées. Une quatrième approche heuristique, l'analyse d'un modèle linéaire non pondéré, est une procédure approximative, mais facile d'emploi. On discute les suppositions et les résultats statistiques de

chaque méthode pour dégager les différences philosophiques de leur origine. On prête attention aux différences de calcul, mais pour aboutir au résultat que des problèmes analogues fournissent des résultats similaires. On pense que les différences essentielles entre méthodes proviennent des suppositions sur lesquelles elles s'appuient et sur la généralité des conclusions qu'elles permettent d'atteindre.

References

Agresti, A. (1980). Generalized odds ratios for ordinal data. *Biometrics* **36,** 59–67.

Amara, I. A. and Koch, G. G. (1980). A macro for multivariate randomization analyses of stratified sample data. *Proceedings of the Fifth Annual SAS Users Group International Conference,* 134–144.

Andersen, E. B. (1980). *Discrete Statistical Models with Social Science Applications.* Amsterdam: North Holland.

Anderson, S., Auquier, A., Hauck, W. W., Oakes, D., Vandaele, W. and Weisberg, H. I. (1980). *Statistical Methods for Comparative Studies.* New York: Wiley.

Andrich, D. (1979). A model for contingency tables having an ordered response classification. *Biometrics* **35,** 403–415.

Baker, R. J. and Nelder, J. A. (1978). *The GLIM System Manual* (*Release* 3). Oxford: Numerical Algorithms Group.

Benard, A. and van Elteren, P. (1953). A generalization of the method of *m* rankings. *Indagationes Mathematicae* **15,** 358–369.

Bennett, B. M. (1968). Rank-order test of linear hypothesis. *Journal of the Royal Statistical Society, Series B* **30,** 483–489.

Berkson, J. (1953). A statistically precise and relatively simple method of estimating the bio-assay with quantal response, based on the logistic function. *Journal of the American Statistical Association* **48,** 565–599.

Bishop, Y. M. M., Fienberg, S. E. and Holland, P. W. (1975). *Discrete Multivariate Analysis: Theory and Practice.* Cambridge, Massachusetts: MIT Press.

Bradley, R. A. (1976). Science, statistics, and paired comparisons. *Biometrics* **32,** 213–233.

Breslow, N. E. and Day, N. E. (1980). *Statistical Methods in Cancer Research; Volume 1—The Analysis of Case Control Studies.* Lyon: International Agency for Research on Cancer.

Clogg, C. C. (1982). Some models for the analysis of association in multi-way cross-classifications having ordered categories. *Journal of the American Statistical Association* **77**. (In the press.)

Cochran, W. G. (1954). Some methods for strengthening the common χ^2 test. *Biometrics* **10,** 417–451.

Cochran, W. G. (1972). Observational studies. In *Statistical Papers in Honor of George W. Snedecor,* T. A. Bancroft (ed.), 77–90. Ames, Iowa: Iowa State University Press.

Cochran, W. G. (1977). *Sampling Techniques.* New York: Wiley.

Conover, W. J. and Iman, R. L. (1981). Rank transformations as a bridge between parametric and non-parametric statistics. *American Statistician* **35,** 124–129.

Conover, W. J. and Iman, R. (1982). Analysis of covariance using the rank transformation. *Biometrics* **38,** 715–724.

Cox, D. R. (1970). *The Analysis of Binary Data.* London: Methuen.

Efron, B. (1971). Forcing a sequential experiment to be balanced. *Biometrika* **58,** 403–417.

Engelman, L. (1981). PLR: Stepwise logistic regression. In *BMDP Statistical Software,* W. J. Dixon *et al.* (eds), Chapter 14.5, 330–344. Los Angeles: University of California Press.

Fienberg, S. E. (1977). *The Analysis of Cross-Classified Categorical Data.* Cambridge, Massachusetts: MIT Press.

Finney, D. J. (1971). *Statistical Method in Biological Assay,* 2nd ed. New York: Hafner.

Friedman, M. (1937). The use of ranks to avoid the assumption of normality implicit in the analysis of variance. *Journal of the American Statistical Association* **32,** 675–699.

Goodman, L. A. (1978). In *Analyzing Qualitative/Categorical Data: Log-Linear Models and Latent Structure Analysis,* J. Magidson (ed.) Cambridge, Massachusetts: Abt Associates.

Goodman, L. A. (1979). Simple models for the analysis of association in cross-classifications having ordered categories. *Journal of the American Statistical Association* **74,** 537–552.

Grizzle, J. E., Starmer, C. F. and Koch, G. G. (1969). The analysis of categorical data by linear models. *Biometrics* **25,** 489–504.

Haberman, S. J. (1974). Log-linear models for frequency tables with ordered classifications. *Biometrics* **30,** 589–600.

Haberman, S. J. (1978). *Analysis of Qualitative Data; Vol.* 1—*Introductory Topics; Vol.* 2—*New Developments.* New York: Academic Press.

Harrell, F. (1980). *SAS Supplemental Library User's Guide,* 1980 *Ed.* Cary, North Carolina: SAS Institute Inc.

Holford, T. R. (1982). Covariance analysis for case–control studies with small blocks. *Biometrics* **38,** 673–683.

Imrey, P. B., Koch, G. G. and Stokes, M. E. (1981, 1982). Categorical data analysis: Some reflections on the log linear model and logistic regression. *International Statistical Review* **49,** 265–283 (Part I); in the press (Part II).

Kempthorne, O. (1955). The randomization theory of experimental inference. *Journal of the American Statistical Association* **50,** 946–967.

Kleinbaum, D. G., Kupper, L. L. and Chambless, L. E. (1982). Logistic regression analysis of epidemiologic data: theory and practice. *Communications in Statistics,* **11,** 485–547.

Koch, G. G., Amara, I. A., Stokes, M. E. and Gillings, D. B. (1980). Some views on parametric and non-parametric analysis for repeated measurements and selected bibliography. *International Statistical Review* **48,** 249–265.

Koch, G. G. and Bhapkar, V. P. (1982). Chi-square tests. In *Encyclopedia of Statistical Sciences,* N. L. Johnson and S. Kotz (eds), 442–457. New York: Wiley.

Koch, G. G., Freeman, D. H., Jr and Freeman, J. L. (1975). Strategies in the multivariate analysis of data from complex surveys. *International Statistical Review* **43,** 59–78.

Koch, G. G. and Gillings, D. B. (1981). Inference, design based vs. model based. In *Encyclopedia of Statistical Sciences,* N. L. Johnson and S. Kotz (eds). New York: Wiley. (In the press.)

Koch, G. G., Gillings, D. B. and Stokes, M. E. (1980). Biostatistical implications of design, sampling, and measurement to health science data analysis. *Annual Review of Public Health* **1,** 163–225.

Koch, G. G., Grizzle, J. E., Semenya, K. and Sen, P. K. (1978). Statistical methods for evaluation of mastitis treatment data. *Journal of Dairy Science* **61,** 830–847.

Koch, G. G., Landis, J. R., Freeman, J. L., Freeman, D. H., Jr and Lehnen, R. G. (1977). A general methodology for the analysis of experiments with repeated measurement of categorical data. *Biometrics* **33,** 133–158.

Koch, G. G. and Stokes, M. E. (1981). Chi-square tests: numerical examples. In *Encyclopedia of Statistical Sciences,* N. L. Johnson and S. Kotz (eds). New York: Wiley. (In the press.)

Kruskal, W. H. and Wallis, W. A. (1953). Use of ranks in one criterion variance analysis. *Journal of the American Statistical Association* **46,** 583–621.

Landis, J. R., Cooper, M. M., Kennedy, T. and Koch, G. G. (1979). A computer program for testing average partial association in three-way contingency tables (PARCAT). *Computer Programs in Biomedicine* **9,** 223–246.

Landis, J. R., Heyman, E. R. and Koch, G. G. (1978). Average partial association in three-way contingency tables: a review and discussion of alternative tests. *International Statistical Review* **46,** 237–254.

Landis, J. R., Stanish, W. M., Freeman, J. L. and Koch, G. G. (1976). A computer program for the generalized chi-square analysis of categorical data using weighted least squares (GENCAT). *Computer Programs in Biomedicine* **6,** 196–231.

Larntz, K. (1978). Small sample comparisons of exact levels for chi-squared goodness of fit statistics, *Journal of the American Statistical Association* **73,** 253–263.

Lehmann, E. L. (1975). *Nonparametrics: Statistical Methods Based on Ranks.* San Francisco: Holden–Day.

Mack, G. A. and Skillings, J. H. (1980). A Friedman-type rank test for main effects in a two-factor ANOVA. *Journal of the American Statistical Association* **75,** 947–951.

Mantel, N. (1963). Chi-square tests with one degree of freedom; extension of the Mantel–Haenszel procedure. *Journal of the American Statistical Association* **58,** 690–700.

Mantel, N. and Haenszel, W. (1959). Statistical aspects of the analysis of data from retrospective studies of disease. *Journal of the National Cancer Institute* **22,** 719–748.

Matts, J. P. and McHugh, R. B. (1978). Analysis of accrual randomized clinical trials with balanced groups in strata. *Journal of Chronic Diseases* **31,** 725–740.

McCullagh, P. (1980). Regression models for ordinal data. *Journal of the Royal Statistical Society, Series B* **42,** 109–142.

Nelder, J. A. and Wedderburn, R. W. M. (1972). Generalized linear models. *Journal of the Royal Statistical Society, Series A* **135,** 370–384.

Neter, J. and Wasserman, W. (1974). *Applied Linear Statistical Models.* Homewood, Illinois: Irwin.

Neyman, J. (1949). Contributions to the theory of the χ^2-test. In *Proceedings of the Berkeley Symposium on Mathematical Statistics and Probability,* J. Neyman (ed.), 239–273. Berkeley: University of California Press.

Pocock, S. J. (1979). Allocation of patients to treatment in clinical trials. *Biometrics* **35,** 183–197.

Puri, M. L. and Sen, P. K. (1971). *Non-Parametric Methods in Multivariate Analysis.* New York: Wiley.

Quade, D. (1967). Rank analysis of covariance. *Journal of the American Statistical Association* **62,** 1187–1200.

Quade, D. (1982). Nonparametric analysis of covariance by matching. *Biometrics* **38,** 597–611.

Semenya, K. A. and Koch, G. G. (1979). Linear models analysis for rank functions of ordinal categorical data. *Proceedings of the Statistical Computing Section of the American Statistical Association,* 271–276.

Semenya, K. A. and Koch, G. G. (1980). Compound function and linear model methods for the multivariate analysis of ordinal categorical data. Institute of Statistics Mimeo Series No. 1323. Chapel Hill: University of North Carolina.

Shirley, E. A. (1981). A distribution-free method for analysis of covariance based on ranked data. *Applied Statistics* **30,** 158–162.

Simon, G. A. (1974). Alternative analyses for the singly-ordered contingency table. *Journal of the American Statistical Association* **69,** 971–976.

Simon, R. (1979). Restricted randomization designs in clinical trials. *Biometrics* **35,** 503–512.

Singer, B. (1979). *Distribution-Free Methods for Non-Parametric Problems: a Classified and Selected Bibliography.* Leicester, England: British Psychological Society.

Snedecor, G. W. and Cochran, W. G. (1980). *Statistical Methods,* 7th ed. Ames: Iowa State University Press.

Stanish, W. M., Gillings, D. B. and Koch, G. G. (1978). An application of multivariate ratio methods for the analysis of a longitudinal clinical trial with missing data. *Biometrics* **34,** 305–317.

van Elteren, P. H. (1960). On the combination of independent two-sample tests of Wilcoxon. *Bulletin of the International Statistical Institute* **37**(3), 351–361.

Wald, A. (1943). Tests of statistical hypotheses concerning several parameters when the number of observations is large. *Transactions of the American Mathematical Society* **54,** 426–482.

Received March 1981; *revised January* 1982