Calculation of Adjusted Response Frequencies Using Least Squares Regression Methods

John E. Overall
University of Texas Health Sciences Center at Houston

The use of general linear regression methods for the analysis of categorical data is recommended. The general linear model analysis of a 0,1 coded response variable produces estimates of the same response probabilities that might otherwise be estimated from frequencies in a multiway contingency table. When factors in the design are correlated, the regression analysis estimates the same response probabilities that would be estimated from the simple marginal frequencies in a balanced orthogonal design. The independent effects that are estimated by the regression analysis are the unweighted means of the response probabilities in various cells of a cross-classification design; however, it is not necessary that all cells in a complex design be filled in order for the estimates to have that interpretation. The advantages of the general linear model analysis include familiarity of most psychologists with the methods, availability of computer programs, and ease of application to problems that are too complex for development of complete multiway contingency tables.

The estimation of response probabilities is one of the most important measurement problems in psychological research. The proportion of subjects producing a "positive" response in a given situation is the relative frequency estimate of response probability. In many situations, however, it is desirable to measure response probabilities independent of a variety of nuisance variables. This article examines the utility of a general linear regression model for estimating response probabilities that would obtain if the effects of several nuisance variables were controlled experimentally. Although the primary concern of this article is with the measurement aspects of the problem, some attention will be paid to tests of significance for differences in estimated response probabilities in different treatment groups.

The analysis of categorical response variables has generally been viewed as a quite different problem from the analysis of normally distributed quantitative variables. Although several contingency table procedures have been proposed for estimating the independent effect of a single factor among several correlated factors (Koch, Johnson, & Tolley, 1972; Mantel, 1963; Shaffer, 1973), they require the construction of complete multiway contingency tables and hence are of restricted utility with regard to practical problems of categorical data analysis involving several control variables. As a consequence, this article examines the pragmatic use of widely available ANOVA and MANOVA computer programs to accomplish analyses of categorical response variables by the same general linear regression methods that can be used to control for numerous nuisance variables in the analysis of quantitative dependent measures.

APPLIED PSYCHOLOGICAL MEASUREMENT
Vol. 4, No. 1 Winter 1980 pp. 65–78
© Copyright 1980 West Publishing Co.

The general linear model regression solution has much to recommend it over contingency table approaches where several partially confounded factors are involved. First and perhaps most important is the fact that it is often impractical to acquire enough data to provide reasonable estimates of response frequencies in the individual cells of a multiway contingency table. With only six independent factors, each involving two or three levels, and one dependent variable having three categories, the total number of cells might be (2×3×3×2×2×2) × 3 = 432. To obtain a sufficient n in each of 432 cells to provide a minimally stable estimate of response probability would seem practically impossible, particularly since some combinations of the independent factors might be difficult to find. The general linear model approach can provide good estimates of the independent effects of six or more factors plus their simple interactions with a feasible sample size, because it is not necessary to fill all cells of the multiway table.

Other reasons why the linear model approach is preferable are (1) computer programs are widely available and familiar to most psychologists; (2) the results to be presented demonstrate adequate robustness; and (3) the adjusted cell proportions can be estimated from the general linear model with only trivial calculations, even though all of the individual cells cannot be estimated directly from the raw data. Limitations on utility of the general linear model for analysis of categorical data include poor fit to rare responses and increasing standard error as the independent factors become highly confounded. General guidelines for excluding such cases will be proposed.

Estimating Response Probabilities as Means of 0,1 Variables

The key to interpretation of general linear model regression estimates of response probabilities is the recognition that the mean of a 0,1 coded response variable is the proportion (or relative frequency) of "positive" responses. Let a positive response be coded 1 and a negative response be coded 0. The sum of the 0,1 coded variable is $\Sigma X = n$, where n is the number of 1 scores. The mean of the 0,1 coded variable is $\Sigma X/N = n/N$, or the proportion of the total N that is coded 1.

Few readers of this journal will be surprised by the mean of a simple 0,1 coded response variable being the proportion of the total sample for whom the response was coded 1. The question of concern in this article is whether the adjusted means for a 0,1 response variable in a nonorthogonal design can properly be interpreted as the response probabilities resulting from the effects of one factor with the effects of other factors held constant. Three types of evidence will be considered to support that interpretation.

First, the equivalence of orthogonal and nonorthogonal designs will be established for the case of a 0,1 coded dependent variable. In a multifactor orthogonal design, the independent effects of each factor on a categorical response variable can be estimated from the marginal response frequencies because the effects of all factors have been experimentally balanced. In a nonorthogonal design, the least squares regression method provides estimates of the same response probabilities that are estimated from the marginal response frequencies in a balanced design involving the same factors. This is an extension of the argument used by Overall, Spiegel, and Cohen (1975) in proposing a criterion for equivalence of orthogonal and nonorthogonal designs for ANOVA with quantitative dependent variables.

The second type of evidence offered in justification of general linear model analysis of 0,1 coded categorical response variables comes from monte carlo studies. The true population response probabilities in various cells of orthogonal and nonorthogonal multiway designs were defined by parameters of a linear model. Categorical response data were generated by sampling from populations having those true response probabilities. The categorical responses were coded 0,1 and entered as dependent variables in a general linear model ANOVA to obtain estimates of the independent effects of each factor on the response probabilities. Across a series of simulations, the results tended to converge on the population response probabilities. The ANOVA tests of significance were found to be nonconservative in the case of nonorthogonal designs, but conservative interpretation should result in valid inferences.

The third type of evidence that is offered in favor of general linear model analysis of categorical data involves the duplicate analysis of real data on alcohol abuse that could be analyzed first by parametric methods and then dichotomized for analysis as a categorical dependent variable. The comparability of results is offered in support of the validity of general linear model analysis of categorical data.

Although the examples to be considered in this article involve 0,1 coded binary response variables, no restriction to binary data is implied. A multicategory response variable can be expanded into *k* different 0,1 binary response variables. That is, a single binary variable can be created to represent membership in each category of a multicategory response variable. The *k* binary response variables can then be analyzed by any of several general linear model ANOVA or MANOVA computer programs that produce the equivalent of a Method 1 analysis as described by Overall and Spiegel (1969). Alternatively, the 0,1 response variables can be entered as dependent variables in a multiple regression analysis, with the independent variables being 0,1,−1 dummy variates resulting from "effects coding" of the various factors whose independent effects on response probabilities are to be evaluated.

Whereas Grizzle, Starmer, & Koch (1969) and Koch et al. (1972) have considered linear model analysis of multiway contingency tables, this article concerns multiple regression analysis of categorical response variables, obviating the need to construct contingency tables. In the simple examples of the next section it will be convenient to display the data in condensed contingency table format; however, such representation becomes unwieldy or impossible as the number of factors multiplies. The general linear model regression analysis handles more complex cases without difficulty.

Comparison of Estimates of Response Probabilities from Orthogonal and Nonorthogonal Designs

The proportion of positive and negative responses in each cell of a multiway contingency table is a relative frequency estimate of the response probability for that combination of factors. The purpose of this section is to confirm that general least squares regression methods can be used to obtain estimates of the independent effect of each factor in a multiway design, irrespective of whether the sample sizes in various cells of the design are equal or unequal. As long as the proportions of positive and negative responses remain constant in the various cells of the design, the estimates of response probabilities obtained from the general linear model analysis of 0,1 coded categorical response variables do not change simply because more subjects are added to some cells of the multifactor design but not to others.

An estimate of the independent effect of each factor in a multiway design is desired. One way to be confident that the effects being examined are due to a particular factor, and not to some other, is to balance the presence of the various factors in the design. For example, an equal number of males and females might be included in each treatment group to ensure that the apparent treatment differences in response probabilities are not actually due to superabundance of female subjects in one of the treatment groups. When it is not feasible to allocate subjects in equal numbers among the various cells of a multiway design, a method of statistical analysis is required that will provide estimates of the same response probabilities that could be estimated in a balanced design involving the same factors. This section demonstrates that least squares regression estimates of response probabilities in a nonorthogonal design are estimates of the same response probabilities that would be obtained from a balanced design involving the same factors. The general linear model analysis estimates the same parameters (response probabilities) in nonorthogonal designs that can be estimated from the simple marginal frequencies in a balanced orthogonal design.

In Table 1 are presented frequencies of success (S) and failure (F) in two multiway contingency tables. In the orthogonal table, there are $N = 10$ subjects in each of the six cells of the A × B design. In the nonorthogonal table, the sample size varies from one cell to the next. The important comparability of the two tables resides in the fact that the *proportions* of S and F responses are the same in comparable cells. The marginal frequencies differ, not because the actual response probabilities differ in any cell, but simply because of differences in the numbers of subjects observed in the various AB treatment combinations.

For analysis by the general linear regression method, the S responses were coded 1 and the F responses were coded 0. The independent factors of the A × B design were coded 0,1,−1 as described in detail by Overall and Spiegel (1969). The complete Method 1 analysis was accomplished as described by those authors. In Table 2 are presented estimates of the effects of Factors A and B on the probabilities of S as calculated from the marginal frequencies in Table 1. For comparison, the adjusted means from the general linear model ANOVA of the 0,1 coded response variable for the nonorthogonal design are shown in the right-hand column. It can be seen that the adjusted means from the nonorthogonal ANOVA are estimates of the same response probabilities that were estimated by the marginal frequencies in the orthogonal design.

For the simple examples of Table 1, equivalent estimates of the independent effects of Factors A and B could be obtained in other ways. The frequencies in each cell of the A × B design could be transformed to proportion of S and the esti-

Table 1
Orthogonal and Nonorthogonal Designs with Equal Response Probabilities in Corresponding Cells

Orthogonal	A_1		A_2		A_3		Row Totals	
	S	F	S	F	S	F	S	F
B_1	2	8	4	6	2	8	8	22
B_2	4	6	8	2	1	9	13	17
Column Totals	6	14	12	8	3	17	21	39

Nonorthogonal	A_1		A_2		A_3		Row Totals	
	S	F	S	F	S	F	S	F
B_1	4	16	8	12	1	4	13	32
B_2	2	3	8	2	2	18	12	23
Column Totals	6	19	16	14	3	22	25	55

Table 2
Estimates of the Independent Effects of Factors A and B on Response Probabilities in Orthogonal and Nonorthogonal Designs

Factors	Marginal Proportions Orthogonal	Marginal Proportions Nonorthogonal	Least Squares Nonorthogonal
A_1	.300	.240	.300
A_2	.600	.533	.600
A_3	.150	.120	.150
B_1	.267	.289	.267
B_2	.433	.343	.433

mates of main effects for Factors A and B calculated as the unweighted means of the cell proportions. The results would be identical for the orthogonal and nonorthogonal cases. The point emphasized here is that the same results *can* be obtained using least squares regression methods. The advantage is that the linear model analysis is much more general than is the complete contingency table approach. With appropriate assumptions concerning absence of higher order interactions, it can be used to partial out the effects of several nuisance variables, even though it would be impractical to consider a complete multiway table in such cases. The general linear model analysis of 0,1 coded response variables and balanced multiway contingency tables yield identical estimates of the independent effects of various factors on response probabilities in simple cases where direct comparison is possible. The inference offered here is that the two approaches should be comparable in more complex situations if it were possible to compare them.

It is important to emphasize that the recommended procedure for estimation of adjusted response probabilities in the analysis of categorical data involves "effects coding" of independent factors and Method 1 strategy, as described by Overall and Spiegel (1969) and Overall, Spiegel, and Cohen (1975). Several computer programs for nonorthogonal ANOVA or MANOVA include such analysis as an option. The BMD 05V, P2V, and 10V (Dixon, 1973), and the SAS (GLM) Type IV (Barr, Goodnight, Sall, & Helwig, 1976) analyses produce Method 1 results. Programs that fit a hierarchy of increasingly more complex models, such as the MANOVA program of Clyde, Cramer, and Shervin (1966) or the MULTIVARIANCE program (Finn, 1976), can be used to obtain appropriate adjusted proportions and tests of significance by successively entering each factor of interest last in the list of independent factors. The SPSS multiple regression program (Nie, Hull, Jenkins, Steinbrenner, & Bent, 1975) can be used to accomplish the analysis by simply coding independent factors 0,1,−1 and the dependent factor 0,1. Thus, numerous standard computer programs exist that can be used to estimate adjusted response probabilities in the analysis of categorical data.

Monte Carlo Studies of ANOVA with Categorical Response Variables

Lunney (1970) accomplished an extensive monte carlo study of the robustness of ANOVA techniques for analysis of 0,1 coded binary responses. The investigation covered a variety of fixed effects ANOVA models with equal numbers of observations in each cell. His conclusion was that ANOVA tests of significance are adequately unbiased if the proportion of responses in the smaller response category is equal to or greater than .2 and there are at least 20 degrees of freedom for error, or if the proportion of responses in the smaller response category is less than .2 but there are at least 40 degrees of freedom for error.

While Lunney's work is important in suggesting robustness of the ANOVA tests on binary response variables, the design of his studies had some shortcomings that limit generalizations to problems of primary concern to the present author. First, the designs he investigated all involved equal numbers of subjects per cell. This meant that one of the chi-square or log linear model techniques for analysis of multiway contingency tables could have been used so that there was no need to force the data into an ANOVA. The concentration on equal cell frequencies leaves unanswered the question of effects of nonorthogonality on the ANOVA tests of binary response variables. In most applications where multiple nuisance variables require statistical control, equal cell frequencies cannot be expected.

As D'Agostino (1972) has pointed out, another primary limitation of the Lunney studies of alpha protection (Type I error rates) was that the same response probabilities were present across all cells in the two- and three-way designs. That is, all effects were null. A major question about the use of ANOVA with binary response variables concerns heterogeneity of variance, since the binomial cell means and variances are p and $p(1-p)$, respectively. It is important to evaluate protection against Type I errors in the testing of some null effects in the presence of other true effects. The presence of true effects will result in pooling of unequal cell variances into the error, even though one or more effects in a complex design may be truly absent. It is important to know whether the Type I error rate remains conservative in that event.

On the positive side, D'Agostino (1972) showed close algebraic similarity between formulae for the usual ANOVA F statistic and the contingency chi-square that is ordinarily considered more appropriate for analysis of frequency data. Moreover, logic suggests that the central limit theorem must hold in the long run, with binomial sampling distributions approaching normal as N increases.

In view of the criticisms that have been raised with regard to previous simulation studies, a series of monte carlo studies was undertaken to evaluate possible bias in parameter estimates and tests of significance in more realistic designs. Stochastic binary response data were generated for a 2^4 factorial design using a random binary generator. The expected binary response frequency in each cell was determined by a linear model that included three nonzero main effect parameters and one simple interaction. All other effects were absent in the model used to generate the data. The linear model that defined the population response probability for each cell in the 2^4 design was of the following form:

$$P_{ijkl} = \mu_0 + \alpha_i + \beta_j + \varepsilon_l + \alpha\beta_{ij}. \qquad [1]$$

The model used to analyze the data included all four main effects, all two-way, and all three-way interactions. Tests of significance were thus provided for 4 effects that were truly present and 10 effects that were truly absent in each analysis. Three hundred analyses were accomplished for balanced (equal cell frequency) designs with n_{ijkl} = 2, 8, and 16. Six hundred analyses were accomplished for a nonorthogonal 2^4 design with cell frequencies of 12 and 4 distributed in such a manner that the phi-coefficient between Factors A and B was $\phi = .50$.

The results from analyses of orthogonal four-way designs involving different sample sizes are presented in Table 3. The true values of the parameters that actually determined response probabilities are shown in the first column. Estimates of those parameters, averaged across runs of 300 analyses are presented in the next three

Table 3

Mean Parameter Estimates and Relative Frequencies of Rejections of Null Hypothesis from 300 Analyses 2^4 Design with Three Sample Sizes

Effect	Parameters	Mean Parameter Estimates			Reject H_0 at α=.05			Reject H_0 at α=.01		
		n=2	n=8	n=16	n=2	n=8	n=16	n=2	n=8	n=16
α_i	.10	.1098	.0964	.1007	.26	.67	.86	.10	.39	.69
β_j	.06	.0534	.0604	.0564	.10	.30	.41	.02	.11	.18
γ_k	.04	.0415	.0454	.0413	.07	.19	.23	.03	.07	.07
ξ_1	.00	-.0017	-.0016	.0060	.03	.06	.02	.01	.01	.00
$\alpha\beta_{ij}$	.10	.0940	.0974	.0990	.20	.65	.86	.07	.40	.66
$\alpha\gamma_{ik}$	.00	-.0019	.0005	.0005	.04	.06	.06	.02	.01	.01
$\alpha\xi_{il}$	.00	.0051	.0003	-.0007	.04	.04	.02	.00	.00	.00
$\beta\gamma_{jk}$	.00	-.0017	-.0015	.0009	.05	.04	.04	.02	.00	.01
$\beta\xi_{jl}$	.00	-.0021	-.0056	-.0007	.04	.06	.03	.01	.01	.01
$\gamma\xi_{kl}$	.00	.0084	.0005	-.0019	.06	.06	.02	.02	.01	.01
$\alpha\beta\gamma_{ijk}$	.00	.0071	-.0015	-.0020	.05	.06	.04	.02	.01	.02
$\alpha\beta\xi_{ijl}$	.00	-.0061	.0046	-.0047	.06	.05	.03	.02	.02	.01
$\alpha\gamma\xi_{ikl}$	.00	.0061	-.0016	.0024	.04	.06	.03	.01	.00	.00
$\beta\gamma\xi_{jkl}$	.00	-.0011	-.0019	.0060	.05	.05	.04	.01	.03	.01
μ_0	.60	.5937	.5937	.5969						

columns. As expected, the complete general linear analysis provided very good estimates of the true effects in the binary response data. The remaining columns of Table 3 indicate the observed proportion of time the null hypothesis was rejected for each effect tested. The extent of bias in alpha protection appears minimal in tests of significance for main effects and interactions that are truly zero in the population. At $\alpha = .05$ the null hypothesis was rejected approximately 5% of the time, and at $\alpha = .01$ the null hypothesis was rejected approximately 1% of the time. Sample size does not appear to influence alpha protection for the analysis of binary response variables in a balanced factorial design.

Next, the same linear model was used to define binary response probabilities in a four-way design, but the cell frequencies were varied to produce a substantial correlation between Factors A and B. Specifically, $n_{ijkl} = 12$ for all cells in which $i = j$, and $n_{ijkl} = 4$ for all cells in which $i \neq j$. As stated earlier, this produced a correlation of $\phi = .50$ between Factors A and B. Results from 600 analyses of a binary response variable in that nonorthogonal design are presented in Table 4. It can be seen that the parameter estimates obtained in analyses of binary response data in this substantially nonorthogonal design converged quite closely on the true population values. Thus, the parameters estimated from analysis of the nonorthogonal design can be recognized to be the same as the parameters estimated in the orthogonal design that involved the same true effects (Table 3). The tests of significance in Table 4 appear somewhat biased in a nonconservative direction; however, the bias does not appear substantial enough to invalidate their use as a basis for statistical inference. With factors no more correlated than $\phi = .50$ or .60, interpretation of the true alpha to be twice the tabled values would appear appropriately conservative.

Numerous other monte carlo simulations of nonorthogonal designs have been accomplished, but the presentation of additional tables of results is perhaps not required. Two other cases should be mentioned, however. It is often said that the most difficult problem from the point of view of regression adjustment in nonorthogonal designs is one in which the unequal cell frequencies correlate highly with the dependent variable. A monte carlo series was run in which the rank correlation between observed cell frequencies and the true response probabilities across 16 cells of a 2^4 design was 1.00. That is, the same linear model that determined the true population response probability in each cell also determined the sample size in that cell. Response probabilities ranged from .40 to .90, and the cell sample sizes ranged from 4 to 9 in perfect rank correlation with the true response probabilities. The results were consistent with the conclusion that an appropriately conservative approach is to simply double the p value for F tests on dichotomous response variables in nonorthogonal designs. Power of the tests of effects which were truly present in this series was somewhat greater than the power represented in results of Table 4 because the variability in cell sample sizes was not as great, even though the correlation with true response probabilities was perfect. The mean parameter estimates, which determine the adjusted response frequencies, were as close in this series as in other cases for which tabular results are presented.

Another difficult case that was examined by monte carlo methods involved a nonorthogonal design with an empty cell. The same 2^4 model was employed with 8 observations in each cell except that $n_{1111} = 16$ and $n_{2222} = 0$. The results from a series of 300 analyses of data generated to have expected response frequencies determined by the same parameters indicated in Tables 3 and 4 revealed near perfect mean estimates of those parameters and nonconservative bias in tests of the null effects. Again, the results were consistent with the recommendation that actual alpha level should be considered double the tabled values when ANOVA F tests are employed as a basis for inferences concerning the significance of treatment effects on response frequencies in nonorthogonal designs. The degree

Table 4
Mean Parameter Estimates and Relative Frequencies of Rejection of the Null Hypothesis From 600 Analyses of a Four-way Nonorthogonal Design

Effect	Population Parameters	Mean Parameter Estimates	Rejection at α=.05	Rejection at α=.01
α_i	.10	.0959	.52	.33
β_j	.06	.0637	.29	.13
γ_k	.04	.0385	.17	.07
ξ_l	.00	-.0010	.06	.02
$\alpha\beta_{ij}$	.10	.0971	.58	.33
$\alpha\gamma_{ik}$	.00	.0016	.06	.02
$\alpha\xi_{il}$	.00	-.0016	.08	.02
$\beta\gamma_{jk}$	.00	-.0009	.06	.02
$\beta\xi_{jl}$	.00	.0018	.09	.02
$\gamma\xi_{kl}$	.00	.0005	.04	.02
$\alpha\beta\gamma_{ijk}$	.00	.0023	.07	.02
$\alpha\beta\xi_{ijl}$	.00	.0035	.08	.02
$\alpha\gamma\xi_{ikl}$	.00	.0026	.06	.02
$\beta\gamma\xi_{jkl}$	.00	.0016	.06	.02
μ_0	.60	.5947		

of bias in the *F* tests is related to the degree of nonorthogonality; so caution is advised where experimental and/or control variables correlate with one another greater than about ϕ = .60; that is the highest degree of nonorthogonality that the present author has examined regarding analysis of categorical response variables.

Effects of Sex and Occupation on Alcohol Abuse

Another way to evaluate the validity of regression adjustment of categorical response variables is to examine data that can be considered appropriate for ordinary analysis of variance, subject it to such analysis, and then arbitrarily categorize it for reanalysis as a binary dependent variable. In this section, a practical problem of some considerable social significance will be examined in the two forms.

Sex differences in the frequency of alcohol abuse have often been reported. A question of fundamental importance is whether the observed differences are biological or socioculturally determined. Basic demographic data were recorded for a sample of 705 patients referred for psychological testing at the University of Texas Medical Branch in Galveston. Alcohol

Table 5
Summary of One-Way ANOVA of Sex Differences Using
4-Point Alcohol Behavior Scale as Dependent Variable

Source	SS	df	MS	F	p
Sex	28.16	1	28.16	28.53	.001
Error	692.66	703	.99		

behavior was recorded in four categories: *abstain, moderate, frequent,* and *problem drinking.* Previous analyses of these data have revealed that the four categories represent an ordinal scale of alcohol abuse.

Recent investigations by Woodward and Overall (1977) have confirmed that 4-point ordered-category scales can be appropriately analyzed by parametric ANOVA methods. Thus, in this report, a general linear analysis of the 4-point ordinal scale data will serve as a meaningful standard against which to compare results from analysis of the same data with the response variable coded in binary form.

In Table 5 is presented the summary of a simple one-way ANOVA undertaken to evaluate sex differences in alcohol abuse using the 4-point scale of alcohol behavior as a dependent variable. A highly significant sex difference in mean alcohol behavior scores is apparent when other factors are disregarded. The summary of a two-way Sex × Occupation ANOVA is presented in Table 6. Those results indicate that the significant *sex difference in alcohol behavior disappears when differences due to occupation are partialed out.* The method of analysis and interpretation were same as discussed by Overall, Spiegel, and Cohen (1975). Because interest here is in illustrating the potential of general linear model analysis of categorical data, no further discussion of the results obtained from analysis of the 4-point alcohol behavior scale will be undertaken at this time.

The alcohol behavior data were next coded 0,1 in two categories. *Frequent* and *problem drinking* were coded 1 to represent alcohol abuse and *abstain* and *moderate* were coded 0 to represent nonabuse. The 0,1 coded dummy variate was entered as the dependent variable in a Method 1 least squares regression analysis, which was in this instance accomplished using the MANOVA program described by Woodward and Overall (1974). That program calculates and prints out adjusted means in the case of nonorthogonal designs, which in this instance can be interpreted as adjusted proportions. The MANOVA program also calculates sums of squares and *F* tests for each dependent variable separately, as well as the multivariate tests of significance. In this analysis, only a single 0,1 dependent variable representing alcohol abuse was analyzed because the proportions of nonabuse are compliments to the proportions of abuse. It should be noted in passing, however, that it is convenient

Table 6
Summary of Two-Way Sex x Occupation ANOVA Using
4-Point Alcohol Behavior Scale as Dependent Variable

Source	SS	df	MS	F	p
Sex	1.46	1	1.46	1.54	N.S.
Occupation	29.52	5	5.90	6.23	.001
Sex x Occup	7.75	5	1.55	1.64	N.S.
Error	656.71	693	.95		

Table 7
Summary of One-Way ANOVA of Sex Differences in
Zero-One Coded Alcohol Abuse Variable

Source	SS	df	MS	F	p
Sex	4.24	1	4.24	21.44	.001
Error	139.02	703	.20		

to code each category of a multicategory response variable as a separate 0,1 variable and to enter the multiple dependent variables into a MANOVA program to obtain directly the adjusted probabilities for all categories of the response variable.

The 0,1 coded alcohol abuse scores were first analyzed in a simple one-way design to calculate the proportions of alcohol abuse for males and females, disregarding other factors. Results from the simple one-way ANOVA of the 0,1 coded response variable are presented in Table 7 for comparison with the results in Table 5. Next, the same binary coded response variable was analyzed in a two-way Sex × Occupation design. This analysis provided adjusted proportions for male and female alcohol abuse with the effects of occupational differences held constant. A summary of the tests of significance from the two-way ANOVA of the 0,1 coded response variable is presented in Table 8 for comparison with the results in Table 6.

The unadjusted proportions of alcohol abuse and nonabuse in the male and female subjects are presented in the left-hand column of Table 9. The adjusted proportions of alcohol abuse, obtained as parameter estimates in the two-way ANOVA of Table 8, are shown on the right in Table 9. It can be seen that statistical adjustment for differences due to occupation (unskilled, skilled, housewife or clerical, sales or business, managerial or professional, and student) tended to reduce the differences between males and females in proportions of alcohol abuse. *The adjusted proportions are interpreted to be estimates of the proportions of alcohol abuse that would be observed for males and females if the samples were carefully equated to contain equal numbers of males and females in all occupation categories.* It is interesting to note that the adjusted proportions differ substantially from the raw proportion for males only. The statistical adjustment had very little effect on the estimated proportion of alcohol abuse in the female population. This would seem to suggest that the often observed sex difference in frequency of alcohol abuse is a sociocultural phenomenon related to occupational role and that it

Table 8
Summary of Two-Way Sex x Occupation ANOVA for
Zero-One Coded Alcohol Abuse Variables

Source	SS	df	MS	F	p
Sex	.055	1	.055	.29	N.S.
Occupation	4.17	5	.83	4.34	.001
Sex x Occup	1.89	5	.38	1.97	N.S.
Error	133.12	693	.19		

Table 9
Raw and Adjusted Proportions of Alcohol Abuse and Non-Abuse for Males and Females Calculated by MANOVA Program

	Raw Unadjusted	Occupation Adjusted
	P	P
Males		
Abuse	.35	.25
Non-Abuse	.65	.75
Females		
Abuse	.19	.20
Non-Abuse	.81	.80

results primarily from the relative rarity of males in occupational roles of homemaker, clerical, and salesperson categories.

The two-way ANOVA of 0,1 coded alcohol abuse suggests that sex differences in alcohol abuse disappear when corrected for occupational level but that occupational level has a significant influence on frequency of alcohol abuse even after sex effects (if any) are partialed out. The raw (unadjusted) and sex-adjusted proportions of alcohol abuse and nonabuse for six occupational groups are presented in Table 10. Contrary to the situation with regard to sex effects, the adjusted proportions for occupational groups do not differ markedly from the unadjusted proportions. This is interpreted as further evidence that occupational role, and not sex per se, is the factor that is primarily responsible for alcohol abuse.

Obviously, it may be questioned whether it is not some broader construct, such as social class, that best accounts for alcohol abuse. However, when this same type of analysis was accomplished using a model that included education and other possible explanatory factors, occupation remained the significant factor and the other social class factors were not found to be important. Age was a factor that was verified to contribute significantly to alcohol abuse independently of occupation. Those results are not reported in detail here because the aim is only to illustrate the potential of a least squares regression approach to analysis of categorical data.

Summary and Conclusions

The primary concern of this article is with the adequacy of least squares regression methods in providing estimates of the independent effects of a single factor on response probabilities, with the effects of several nuisance variables held constant. It was demonstrated that the general least squares regression analysis of a 0,1 coded response variable produced estimates of independent effects without requiring independence among factors in a multiway design. Specifically, the regression estimates in a nonorthogonal design estimate the same effects that might otherwise be estimated from the marginal response frequencies in a balanced orthogonal design involving the same factors.

Monte carlo methods were employed to generate binary responses having known true population probabilities in various cells of 2^4 factorial designs in which some effects were truly present and some were truly absent. Designs involving equal cell frequencies were considered first, and then designs involving unequal cell frequencies were examined. In all cases, the regression estimates of the effects of various factors converged toward the known population values across a series of several hundred simulated experiments.

Table 10
Raw and Adjusted Proportions of Alcohol Abuse and Non-Abuse for Six Occupational Groups Calculated by MANOVA Program

	Raw Unadjusted	Sex Adjusted
Unskilled		
Abuse	.28	.27
Non-Abuse	.72	.73
Skilled		
Abuse	.35	.31
Non-Abuse	.65	.69
Housewife/Clerical		
Abuse	.15	.08
Non-Abuse	.85	.92
Sales/Business		
Abuse	.16	.12
Non-Abuse	.84	.88
Managerial/Professional		
Abuse	.53	.58
Non-Abuse	.47	.42
Student		
Abuse	.00	.00
Non-Abuse	1.00	1.00

The ANOVA *F* tests appeared adequately unbiased in designs involving equal cell frequencies; however, the ANOVA tests on binary response variables tended to evidence a nonconservative bias when sample sizes in various cells of the multiway designs differed substantially. It is recommended that the use of a more stringent alpha level can provide appropriate alpha protection in the analysis of binary coded response variables in nonorthogonal designs.

In a final example, real data concerning alcohol abuse in male and female psychiatric patients was first analyzed as a 4-point scale and then as a 0,1 coded binary response variable. The comparability of results further confirmed validity of the general linear model analysis of binary coded data for the purpose of partialing out the independent effects of related factors. Disregarding other factors, the proportion of male patients with histories of alcohol abuse is greater than the proportion of female patients with such histories. However, when sex and occupation were entered in a two-way general linear model ANOVA, with alcohol abuse coded as a 0,1 dependent variable, the differences in occupation-adjusted proportions of alcohol abuse between the male and female samples were substantially reduced from the differences observed without adjusting for occupation. Significant differences between occupational groups in relative frequencies of alcohol abuse remained after the sex effect was partialed out. The results suggest that it is the differences in social roles and environment (as reflected in occupation) that accounts for the greater frequency of alcohol abuse among males. The greater frequency of alcohol abuse in certain occupational groups is apparently not because those occupations are pursued primarily by men.

The general linear model analysis of 0,1 coded categorical data appears to offer a useful alternative to multiway contingency table analysis. Where contingency table analysis is feasible, the general linear model analysis provides equivalent results. It is generally impractical to consider the contingency table approach where several partially correlated factors are involved. Psychologists and others who are familiar with computer programs for general linear model ANOVA or least squares multiple regression will find those methods meaningful and convenient for estimation of adjusted response probabilities without resort to complex contingency table analyses.

References

Barr, A. J., Goodnight, J. H., Sall, J. P., & Helwig, J. T. *A user's guide to the statistical analysis system '76.* Raleigh, NC: SAS Institute, Inc., 1976.

Clyde, D. J., Cramer, E. M., & Shervin, R. J. *Multivariate statistical programs.* Coral Gables, FL: University of Miami, Biometric Laboratory, 1966.

D'Agostino, R. B. Second look at analysis of variance on dichotomous data. *Journal of Educational Measurement,* 1971, *8,* 327-330.

D'Agostino, R. B. Relation between the chi-squared and ANOVA tests for testing the equality of *k* independent dichotomous populations. *American Statistician,* 1972, *26,* 30-32.

Dixon, W. J. (Ed.), *BMD: Biomedical computer programs.* Los Angeles: University of California Press, 1973.

Finn, J. D. *MULTIVARIANCE: Univariate and multivariate analysis of variance.* Ann Arbor, MI: National Educational Resources, 1978.

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

Koch, G. G., Johnson, W. D., & Tolley, H. D. A linear models approach to the analysis of survival and extent of disease in multi-dimensional contingency tables. *Journal of the American Statistical Association,* 1972, *67,* 783-796.

Lunney, G. H. Using analysis of variance with a dichotomous dependent variable: An empirical study. *Journal of Educational Measurement,* 1970, 7, 263-269.

Mantel, N. Chi-square tests with one degree of freedom; extensions of the Mantel-Haenszel procedure. *Journal of the American Statistical Society,* 1963, *58,* 690-700.

Nie, N. H., Hull, C. H., Jenkins, J. G., Steinbrenner, K., & Bent, D. H. *Statistical package for the social sciences* (2nd ed.). New York: McGraw-Hill, 1975.

Overall, J. E., & Spiegel, D. K. Concerning least squares analysis of experimental data. *Psychological Bulletin,* 1969, *72,* 311-322.

Overall, J. E., Spiegel, D. K., & Cohen, J. Equivalence of orthogonal and non-orthogonal analysis of variance. *Psychological Bulletin,* 1975, *82,* 182-186.

Shaffer, J. P. Defining and testing hypothesis in multi-dimensional contingency tables. *Psychological Bulletin,* 1973, *79,* 127-141.

Woodward, J. A., & Overall, J. E. A general multivariate analysis of variance computer program. *Educational and Psychological Measurement,* 1974, *34,* 653-662.

Woodward, J. A., & Overall, J. E. The significance of treatment effects in ordered category data. *Journal of Psychiatric Research,* 1977, *13,* 169-177.

Acknowledgments

This work was supported in part by grant DHEW-1-R01-MH32457-01.

Author's Address

Send requests for reprints or further information to John E. Overall, Department of Psychiatry, University of Texas Health Sciences Center at Houston, Medical School, P. O. Box 20708, Houston, TX 77025.